Full-School Engagement as a Mediator of Ethnic and Economic Composition Effects on Grade 8 Mathematics Test Scores: A Two-level Structural Equation Model

Tom Munk

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Education

Chapel Hill
2007

Approved by

Professor Susan N. Friel

Professor Kenneth Bollen

Associate Professor Carol Malloy

Assistant Professor James Trier

Professor William B. Ware

ABSTRACT

TOM MUNK: Full-School Engagement as a Mediator of Ethnic and Economic Composition Effects on Grade 8 Mathematics Test Scores: A Two-level Structural Equation Model
(Under the direction of Susan N. Friel)

The purpose of this study is to quantitatively investigate one of many possible reasons for gaps in grade 8 students' mathematics test scores between students of different ethnicities or economic levels. Recent advances in multi-level structural equation modeling, together with increased sample sizes available from the National Assessment of Educational Progress (NAEP, also known as The Nation's Report Card) allow for a renewed investigation of the factors that explain ethnic or economic test score gaps in the United States. Using preliminary and confirmatory samples from the 2003 grade 8 NAEP mathematics assessment and survey results, the current study estimates a basic two-level model of ethnic and economic predictors of mathematics test scores and explores one possible mediator of ethnic and economic composition effects.

Within schools, the study confirms previous studies documenting that White, Asian, or higher income students tend to score higher than lower-income, Black, Hispanic, or American Indian students. Between schools, the model suggests that schools with higher percentages of lower-income students are less effective for all of their students than schools with higher percentages of higher-income students. No such composition effects are confirmed based on ethnic composition, that is, the percentages of Black, Asian, or American

Indian students in a school. Unexpected composition effects are found suggesting that schools with higher percentages of Hispanic students and schools with higher percentages of Title I students are more effective for all of their students than schools with lower percentages in these categories. Effective Title I funding, social capital in the Hispanic community, and effective school response to large numbers of Spanish-speaking students are suggested as explanations.

A successful confirmatory factor analysis is performed on one potential mediator of composition effects – a second-order construct called Full-School Engagement (FSE). FSE is shown to be a partial mediator of the effect of school economic composition on grade eight adjusted mean mathematics test scores. No other composition effects are consistently mediated by FSE. This study demonstrates a successful application of two-level structural equation modeling using the rich, but complex, NAEP database.

ACKNOWLEDGEMENTS

I would like to thank:

- Kenneth Bollen; I was always amazed by the responsiveness of "the man who wrote the book" to my questions about structural equation modeling and by the clarity and kindness of his responses;
- Pierre Bourdieu, whose book, *Reproduction*, provides a frame for this and, I hope, many future studies;
- The Chapel Hill Monthly Meeting of the Religious Society of Friends, for almost two decades of love and support;
- Sharon Christ for getting me started with Mplus;
- Gregory Cizek for introducing me to NAEP, encouraging me to be trained in its use, and helping me to get access to the data;
- John Dossey, for his early advice and support;
- Tom Fiore, for believing in me and giving me the space I needed to complete this work;
- Susan Friel, my advisor, for working patiently with me as I struggled for written clarity;
- Mark Gould, for teaching me to analyze how power, money, and ideology distort our democratic ideals;

- Rodney Hodson for his help and support in finding me the computer and the secure space I needed to use the NAEP data; also for being the friendly face and voice in that space;
- Jesus of Nazareth, for showing us the power of Love;
- Stas Kolenikov; the generous support and friendship of this statistician was invaluable;
- Jennifer Leeman, for everything; a hundred pages would not be enough;
- Sarah Lubienski for showing me the way with her example and her support on multiple occasions and in multiple domains;
- Sam and Rachel Leeman-Munk, for growing up so well despite living with a father who was frequently absent in body and even more frequently absent in mind;
- Carol Malloy for her insistence that I understand issues of class and race deeply before I put pen to paper;
- Karl Marx, for his example of courageous and uncompromising dedication to the use of science to speak truth in support of equality;
- Joyce Munk for making me believe from the moment I was born to this day that I am safe and loved;
- Paul Munk for showing me how to stay humble, question everything, and think hard;
- Linda Muthén for her incredibly responsive support of the outstanding Mplus product;

- Andreas Oranje for helping me to design my responses to the challenges of research with NAEP data;
- The people of the state of North Carolina for funding the University that provided my intellectual support system;
- Carolyn and Thomas Royster for generously sponsoring the fellowship program that made my PhD possible;
- Alex Sedlacek for her leadership in support of NAEP research and her kindness to me;
- Lynda Stone for her unfailing friendliness and invaluable help in obtaining the Royster Fellowship;
- Jim Trier for graciously filling a committee hole at the last minute, for providing an example of successful counter-hegemonic scholarship, and for teaching the best course I took in my doctoral program;
- Bill Ware, for introducing me with his marvelous lectures to the thrill of statistical analysis;
- Harold Wenglinsky, for showing me that multilevel structural equation modeling of NAEP data could be done, and for his responsiveness to my questions about his work.

The strengths of this dissertation belong to them as much as to me. The weaknesses are mine alone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE - INTRODUCTION

The No Child Left Behind Act of 2001 (NCLB) requires states, districts, and schools to bring all students, in all major subgroups, to the level of proficiency or above in mathematics by the 2013-2014 academic year (Gingerich, 2003, p. 12; Kim & Sunderman, 2005). According to the 2005 National Assessment of Educational Progress (NAEP), mathematics proficiency is much less common among poor[1] eighth graders (13%) than their non-poor (39%) peers. Similar gaps exist between Black (9%), Hispanic (13%), and American Indian (14%) eighth graders and their White (39%) or Asian (47%) peers (National Center for Education Statistics, 2005). Similar findings have been demonstrated in numerous studies (Coleman et al., 1966; Jencks & Phillips, 1998; Strutchens, Lubienski, McGraw, & Westbrook, 2004; White, 1982). In an equitable society, such disparities in mathematics test scores[2] associated with ethnicity or economic level would not occur (Gutiérrez, 2002). To achieve even the limited definition of equity proposed by NCLB, we must solve the puzzle of why students who are poor or members of certain ethnic subgroups continue to score lower on standardized mathematics tests than their peers.

---

[1] The primary indicator of economic level for NAEP and NCLB is free or reduced-price lunch eligibility.

[2] Many researchers use the terms "mathematics test scores" and "mathematics achievement" interchangeably. Following the example of Jencks and Phillips (1998), this study uses "test scores" in order to avoid the faulty presumption (Rogoff & Chavajay, 1995) that test scores are accurate, complete, and culturally neutral measures of mathematics achievement.

In the U.S. system, economic and ethnic segregation between schools[3] may tend to place poor students and students from some ethnic subgroups in schools that are less effective in promoting achievement (Cashin, 2004; Kozol, 1991, 2005; Orfield, 1996). The school provides a context for learning. Given the economic and ethnic divisions apparent in our society, the economic or ethnic makeup of the school is a salient part of that context. Some studies have suggested that the economic or ethnic composition of a school may impact the achievement of all students in that school (Bankston III & Caldas, 1996; Myers, 1985; Schellenberg, 1999). These relationships are referred to as *composition effects* (Raudenbush & Bryk, 2002; Willms, 2006). Composition effects cannot help explain how the test scores of different groups of students vary within schools; rather, they focus attention on how and why test scores vary between schools.

The literature offers five reasons why economic or ethnic composition effects may exist. Specifically, schools serving larger proportions of students who are poor or members of certain ethnic subgroups may (a) need to deal with a wide range of needs, including educational, emotional, physical, and medical needs (Rothstein, 2004); (b) experience a lack of the resources money can buy (Darling-Hammond, 2004; Grissmer, Flanagan, Kawata, & Williamson, 2000; L. Harris, 2004; Kozol, 1991; Strutchens et al., 2004); (c) face staffing problems, including recruitment and retention of quality teachers (Betts, Rueben, & Danenberg, 2000; R. F. Ferguson, 1991; Goe, 2002); (d) experience dissonance between the dominant culture that pervades schooling and the cultures that students bring from their homes and communities (Berry III, 2002; Bourdieu & Passeron, 1977; Delpit, 1995; Malloy

[3] Economic and ethnic segregation also occur within schools (i.e. via tracking). This within-school segregation may increase the size of test-score gaps (Oakes, Johnson, & Muir, 2004), but that analysis is not a part of this study.

& Malloy, 1998; Rogoff, 2003, p. 85); or (e) experience lower engagement levels of students (Fordham & Ogbu, 1986; P. Willis, 1981), parents (Martin, 2000), teachers (R. F. Ferguson, 1998; Rist, 1970), and administrators (M. M. Harris & Willomer, 1998). The first three reasons are more easily quantified and, thus, measured and reported. The last two reasons are less easily quantified and, as a consequence, are difficult to measure.

This focus of this study is on the last of these reasons – engagement. A new construct, Full-School Engagement, is proposed. It is defined as the degree to which an entire school community – students, parents, teachers, and administrators – is actively engaged in the academic mission of the school. Many researchers have focused on the engagement of one or more of these subgroups in the schooling process, but research has not adequately addressed Full-School Engagement, a property of a school that incorporates the engagement of each of these subgroups. Some literature exists (Comer, Haynes, Joyner, & Ben-Avie, 1996; V. E. Lee, Bryk, & Smith, 1993; Raudenbush, Fotiu, & Cheong, 1998; Wenglinsky, 1997) to suggest that constructs similar to Full-School Engagement are a part of the reason for economic or ethnic composition effects, but there has been no large-scale quantitative analysis bringing the full construct together and measuring its impact on ethnic and economic test score gaps.

Model and Questions

The purpose of this study is to investigate the degree to which Full-School Engagement explains the grade 8 mathematics test score gaps that exist between economically or ethnically differing groups of students. This investigation addresses the following four questions:

Question 1. Can a single second-order latent variable called Full-School Engagement measure a constellation of factors representing administrative, parent, teacher, and student engagement in the academic mission of a school?

Question 2. Does the economic or ethnic composition of a school predict that school's mean grade 8 mathematics tests scores, adjusted for the ethnicity and economic level of the individual students in that school (i.e., composition effects)?

Question 3. What are the relationships that exist among the economic and ethnic compositions of a school, Full-School Engagement, and adjusted school mean grade 8 mathematics test scores?

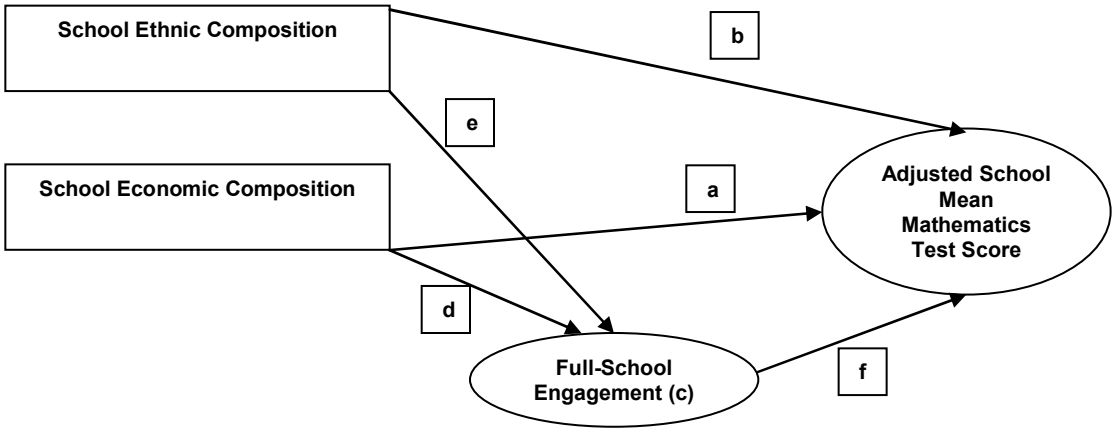
Question 4. Does Full-School Engagement mediate any of the composition effects identified in Question 2?

Figure 1 is a path model of the proposed relationships among Full-School Engagement, school economic and ethnic compositions, and grade 8 adjusted school mean mathematics test scores. The model proposes that the economic and ethnic compositions of a school each affect that school's mathematics test scores directly (paths *a* and *b*). The same economic and ethnic compositions affect the level of Full-School Engagement (paths *d* and *e*), creating an indirect effect on scores (paths *df* and *ef*). All of these composition[4] effects are hypothesized to remain significant even controlling for the within-school effects of individual student economic level (i.e., student lunch status, student Title I status) and ethnicity on mathematics

[4] As described in greater detail in the methodology section, within-school effects answer questions about why certain students within a given school earn higher test scores than other students. Composition effects explain why certain schools generate higher test scores than others, controlling for all within-school variance. Between-school effects are the sum of within-school and composition effects. (Raudenbush & Bryk, 2002; Willms, 2006).

test scores.[5] As described briefly here and in more detail in chapter 2, the various components and relationships built into this model are supported with theory and research. However, the model has not been tested as a whole in a large-scale quantitative study. The present study proposes to address this gap.

Figure 1. Full-School Engagement Mediation Model



The National Assessment of Educational Progress

The 2003 National Assessment of Educational Progress (NAEP) grade 8 mathematics assessment provides a tool to test this model. This assessment includes data from a very large sample of 153,000 eighth graders from 6,100 schools (National Center for Education Statistics, 2003, p. 20). It includes a large collection of background variables, several of which provide good measures for parent, student, teacher, and administrative engagement (National Center for Education Statistics, 2004). It is designed to align with the influential *Standards* of the National Council of Teachers of Mathematics (Lindquist, 2001; National

[5] The within-school controls model is not pictured in the diagram, but it is these controls for student ethnicity and economic level that effectively adjust the school mean mathematics scores for the ethnicity and economic levels of the students in the school.

Council of Teachers of Mathematics, 1989, 2000). NAEP is called "The Nation's Report Card" because it has been designated in the No Child Left Behind legislation as the tool to be used for validation of state test results (Miller, 2003). The test is administered to fourth, eighth, and twelfth grade students in the spring. The eighth-grade sample is used because it has less missing data than the twelfth-grade sample and because there is greater reliability in the background information than may be found with the fourth-grade sample.[6]

The hypotheses of this study have been suggested by prior research. The contributions of this study are to provide a quantitative test of the significance and magnitude of economic and ethnic composition effects, to broaden the definition of school engagement with a well-measured construct called Full-School Engagement, and finally, to investigate whether Full-School Engagement mediates the effects of school economic or ethnic compositions on adjusted school mean mathematics test scores.

Overview of Chapters

Chapter Two places the current study in the context of prior research, providing justification for the questions and the hypothesized model. Chapter Three describes the NAEP data and specifies five increasingly complex models to be used in the study, along with methods and data used to analyze each model. Chapter Four describes the results of analyses of each of the five proposed models. Chapter Five summarizes, evaluates, and interprets the results presented in Chapter Four with respect to the original research questions and the proposed model. Theoretical and practical consequences of the study, the validity of

[6] Older students are more likely to respond accurately to questions about their home backgrounds than younger children (Grissmer et al., 2000).

the conclusions, limitations of the study, and suggestions for future work are included in this final chapter.

CHAPTER TWO - LITERATURE REVIEW

The purpose of this study is to quantitatively investigate one of many possible reasons for gaps in grade 8 students' mathematics test scores between students of different ethnicities or economic levels. The hypothesis to be investigated is:

1. the ethnic or economic composition of a school's student body predicts student, teacher, and parent engagement in the schooling process (Full-School Engagement);
2. in all categories of schools, increased Full-School Engagement leads to improved mathematics test scores; and
3. therefore, Full-School Engagement partially explains ethnically and economically based gaps in grade 8 mathematics test scores.

This chapter systematically reviews the literature surrounding this hypothesis. As an organizing preface to the full literature review, the complete basic argument is presented in brief without citations.

The Argument in Brief

A student's mathematics test scores can be partially predicted by the ethnicity or economic level[7] of that student. Ethnic test score gaps cannot be fully explained by socioeconomic

[7] The economic level of a student can reasonably be seen as equivalent to the economic level of that student's parents. The economic level of the parents is measured in various ways by various researchers. It is generally conceived as a continuous variable ranging from low income to high income. In the design for this study, economic level is measured by students' free or reduced lunch status and by students' eligibility for Title I.

differences. Neither can economic test score gaps be explained by ethnic differences. Both ethnicity and economic level are essential pieces of any explanation of ethnic and economic mathematics test score gaps. Within any given economic level, members of ethnic groups considered to be involuntary minorities[8] tend to have lower mathematics test scores than their majority and voluntary minority counterparts. Within any ethnic group, higher-income students tend to have higher test scores than lower-income students.

Mathematics test score gaps based on student ethnicity and economic level occur both within schools and between schools. Within any given school, the ethnic and economic mathematics test score patterns described above usually hold true. However, the ethnic or economic composition of a school's student body may predict a school's effectiveness,[9] which further contributes to test score gaps. Continuing ethnic and economic segregation of schools may mean that involuntary minority students and students from poorer families tend to go to schools that are less effective for *all* of the students in those schools. These are called *composition effects*.

These composition effects may occur for five interrelated reasons.

---

[8] The majority and dominant culture in the U.S. is European-American, or White. Asians are considered to be a voluntary minority group because the vast majorities have come to this nation in search of greater opportunity. Black communities mostly arrived as slaves, and Indians were incorporated through a process of conquest, displacement, and cultural genocide. These are considered involuntary minority groups. Hispanics are a semi-voluntary group because, while many have arrived in the U.S. voluntarily, the original Spanish speaking group was incorporated via military conquest. Worldwide, majority and voluntary minority groups perform better in schooling systems than students from groups that joined a nation involuntarily (Ogbu, 1978, 1988, 1992, 1997; Ogbu & Simons, 1994).

[9] In this study, schools that produce higher test scores, even controlling for the backgrounds of the students in the schools, are referred to as "effective schools." There is a large body of research, called effective schools research, investigating the characteristics of such schools (American Association of School Administrators, 1992; Cohen, Raudenbush, & Ball, 2003; Hawley, 2002; V. E. Lee et al., 1993; Levine & Lezotte, 1995; Mullis, Jenkins, & Johnson, 1994; Newmann, 1992; Patchen, 2004).

1) Students may come to school with greater and more varied needs, necessarily diverting attention from the academic mission of the school.

2) There may be less financial resources, leading to a lower quantity and quality of facilities and instructional materials.

3) Attracting and retaining high-quality staff may be more difficult. This may be related to weaker curricular offerings and less emphasis on reasoning.

4) Cultural dissonance may lead to conflict within the school.

5) All parties involved in the schooling process may engage less in the academic mission of the school.

The association between composition effects and lack of engagement, described in (5) above, may occur across numerous parties involved in the schooling process. Teachers are more engaged in their work when schools have more high-income or less involuntary minority students (Teacher Engagement). Parents of higher-income, majority, or voluntary minority students are more likely to be engaged in the schooling process, improving the learning and test scores of all the students in the schools their children attend (Parent Engagement). In schools with many lower-income or involuntary minority students, school cultures may favor less engagement in the schooling process, leading to lower test scores (Student Engagement). Adolescent lower-income students or involuntary minority students may be more likely to actively disrupt the learning environment in a school (Student Resistance). Schools with fewer lower-income or involuntary minority students may have more optimistic leadership. This may help make such schools more effective (Administrative Optimism). The engagements of parents, teachers, students, and administrators are so highly

related that they might be most usefully viewed as dimensions of a single construct: Full-School Engagement.

Because of these specific relationships, Full-School Engagement may be part of the explanation for the effects of school ethnic or economic composition on school effectiveness. In statistical terms, it may *mediate* those relationships. The primary purpose of this study is to investigate this relationship using a large, nationally representative database

The 2003 Grade 8 Mathematics National Assessment of Educational Progress (NAEP) provides an ideal data source for this study. NAEP provides a very large sample size and is the only nationally representative survey of what students know and can do in mathematics and other school subjects. The test is designed to measure conceptual understanding, procedural knowledge, and problem solving in five mathematics content areas and is designed with the highest psychometric standards. NAEP also surveys teachers and administrators of the tested students. The administrator survey provides a good set of data for the investigation of the Full-School Engagement construct. NAEP is moving in the direction of becoming a de facto national test, rapidly increasing its importance to students of all backgrounds.

Section 1. Test Score Gaps

The Coleman Report

This study is an example of an education production function (Burtless, 1996; Greenwald, Hedges, & Laine, 1996; Hanushek, 1986, 1996a, 1996b; Hedges & Greenwald, 1996; Jefferson, 2005; Levačić & Vignoles, 2002; Pritchett & Filmer, 1999; Wenglinsky, 1997). The purpose of such studies is to quantitatively investigate the effects of various resources on

student test scores. The focus of this study is on a resource called Full-School Engagement, but many other resources (e.g., teacher quality, lab space, libraries, student-teacher ratios) have been investigated. The seminal education production function study, *Equality of Educational Opportunity* (Coleman et al., 1966), was commissioned by Congress as part of the Civil Rights Act of 1964 and is often referred to as the Coleman Report.

Coleman and his fellow researchers were charged by Congress to investigate inequality of educational resources among various ethnic groups in the nation's schools. Their surveys and tests included 570,000 students and 60,000 teachers in 4,000 elementary and secondary schools across the country. The Coleman Report included many sub-studies; these are referred to in the following paragraphs as Coleman studies.

Although they were commissioned to investigate school resource inputs, Coleman and his colleagues also associated those inputs with educational outputs in the form of test scores. Borrowing advanced regression techniques and a single-outcome focus[10] from economists, they performed the seminal education production function study. Their results were shocking to those who pinned their hopes for equal opportunity on an equalization of school resources. The researchers reported that characteristics of a student's family and of a student's community influenced test scores far more than the school resource differences they measured.

The field of education production function research has been productive and contentious since the time of the Coleman Report. Although some have found that money spent on schools buys resources that have an effect on students' test scores (Dewey, Husted, &

[10] For years, economists had used advanced regression techniques to analyze the best ways to make money. These had been called production function studies. Coleman borrowed the regression techniques and picked an educational analogue to money – test scores – as the sole outcome variable for his study.

Kenney, 2000; Figlio, 1999; Goldhaber & Brewer, 1997; Greenwald et al., 1996; Jefferson, 2005; Levačić & Vignoles, 2002; Wenglinsky, 2002, 2004); others have found that the resources on which most education dollars are spent have little substantial impact on test performance (Hanushek, 1986, 1996a; Pritchett & Filmer, 1999; Xin, Xu, & Tatsuoka, 2004). None of the researchers, however, disagree with the claims that student background and school composition are among the important predictors of school success.

The Coleman Report's analysis was detailed and statistically sophisticated. It defined eight demographic groups in language that is now dated: Mexican Americans, Puerto Ricans, Indian Americans, Oriental Americans, Southern Negroes, Northern Negroes, Southern Whites, and Northern Whites. Each group was analyzed separately; the relative importance of various test score predictors was compared between groups. The primary outcome variable was a test of verbal ability.[11] Unlike the majority of studies both before and after, the Coleman Report took care to statistically separate within-school effects from between-school effects.

The focus of the Coleman Report was the influence of schools on achievement, but the researchers controlled for student background characteristics because these characteristics "shape the child before he reaches school" (p. 298). They surveyed secondary students about objective background factors (i.e., urbanism of their background, parents' education, the

[11] The test of verbal ability was chosen because it showed the highest correlation with other test scores and was at least as much affected by school differences as scores on a variety of achievement tests. The researchers suggested that "the similarities between schools tend to compress the school-to-school component of variance in subjects toward which the curriculum is directed; the differences between schools became evident in the things their students learn, covered in ability tests, that are not as directly related to the curriculum" (Coleman et al., 1966, p. 294). Since mathematics tends to be taught quite directly and similarly in the vast majority of schools, the between-school variation in reading comprehension scores might be expected to be larger than in mathematics achievement scores. This is what Coleman et al. found.

presence of a mother and father in the home, family size, consumer items, and reading material in the home) and subjective background factors (i.e., parental interest in education and parental desires for educational accomplishment). Taken together, these eight background factors explained between 6% (Puerto Ricans) and 23% (Northern Whites) of the total variance in grade nine verbal test scores.

For each demographic group, total test score variation was divided into two parts: (1) the variations of individual pupils' scores from the mean score of their ethnic group in the school (within-school variance); and (2) variations of school means about the group's mean score in the nation (between-school variance) (Coleman et al., 1966, p. 296). Using verbal ability test scores at grades 1, 3, 6, 9, and 12 for the eight groups, Coleman et al. found that between 5% (Oriental Americans, grade 12) and 38% (Indian Americans, grade 3) of total test score variance happened between schools. Although most of the variance (62% to 95%) was within schools, the between-school portion (also called an Intra-Class Coefficient or ICC) was large enough to warrant a two-level (within-school and between-school) analysis.

The between-school differences explained approximately the same share of test score variance as the background measures. The researchers noted, however, that the background variable probably underestimated the importance of family background because the addition of more background variables to the model could increase the variance explained. On the other hand, the between-school differences in test scores overestimate differences in school quality because between-school differences can also be caused by "differences from one community to another in family backgrounds of individuals including abilities of students" or "differences from one community to another in influences outside school, apart from the student's own family" (Coleman et al., 1966, p. 295). Because between-school test score

differences increased significantly across grades for only two of the eight demographic groups,[12] the researchers concluded that "the larger part of school-to-school variation in achievement appears to be not a consequence of effects of school variations at all, but of variations in family backgrounds of the entering student bodies."

Nevertheless, between-school test score differences were far from negligible, and the researchers wanted to explain them. Here again, they declined to point the finger in the expected "resources" direction. They found student body composition to be far more influential. "Attributes of other students account for far more variation in the achievement of minority group children than do any attributes of school facilities and slightly more than do attributes of staff" (p. 302). Specifically, they controlled for the background factors of individual students, and found that overall student body aspirations and educational background[13] were far more strongly related to school test scores than a set of eleven school characteristics.[14]

They also looked at school ethnic composition by adding a "proportion White" variable to their models. They found that schools with more White students had better performance by students in all racial categories and that this effect was not explained by school resource

---

[12] Between grade 1 and grade 12, the percentage of total variance in individual verbal achievement scores lying between schools rose from 17% to 22% for Puerto Ricans and from 19% to 31% for Indian Americans. It fell for the other six groups.

[13] The student-body characteristics were: proportion whose families own encyclopedias, number of student transfers, attendance, proportion planning to attend college (grades 9 and 12 only), teachers' perception of student-body quality (1, 3, 6 only), and average hours of homework (9 and 12 only).

[14] The school characteristics included in the grade 9 analysis were: per-pupil expenditure on staff, volumes per student in library, science lab facilities, extracurricular activities, presence of accelerated curriculum, comprehensiveness of curriculum, use of tracking, movement between tracks, size, guidance counselors, and school location.

differences, but was largely explained by "the better educational background and higher educational aspirations that are, on the average found among White students" (p. 307).

The pioneering Coleman Report had limitations, but it raised important issues about the relative importance of various physical and human resources, both at the student level and at the school and community levels, in the generation of inequality of educational outcomes. More recent studies are addressed below to investigate the relationships of ethnicity and economic level with test scores, particularly mathematics test scores.

Student Economic Level and Test Scores

The Coleman Report found a strong relationship between a student's background and her verbal test score, with a weak measure of family economic level providing only a small piece of that explanation. Two subsequent meta-analyses of studies of socioeconomic status and test scores (Sirin, 2005; White, 1982), including 175 studies altogether, found somewhat stronger effects. Operationalizations of socioeconomic status (SES) and achievement made a difference to the results. Choice of unit of analysis also had a large effect on the results; few in the four decades since the Coleman Report have handled the unit of analysis problem as well as Coleman et al. A discussion of these three studies follows.

In the Coleman Report, six objective background factors (i.e., urbanism, parents' education, structural integrity of home, smallness of family, consumer items in home, reading materials in home) explained between 4% (Puerto Rican) and 18% (Southern White) of grade 9 verbal ability test score variance. Adding subjective background factors (i.e., parent's interest and educational desires) to the model explained an additional 1% (Puerto Rican) to 10% (Oriental) of the variance (Coleman et al., 1966, table 3.221.3, p. 300). Controlling for the seven other variables, the economic level of the family (as measured by a survey of

consumer items in the home) explained less than 1% of the variance for each ethnic group at grade 9 (Coleman et al., 1966, table 3.221.6, p. 301).

White (1982) and Sirin (2005) both noted that although SES, broadly defined, has been shown to be strongly related to all kinds of achievement variables, there is little consistency across studies in the theoretical understanding of what Coleman et al. vaguely called background variables and what subsequent researchers have usually called SES. White's (1982) meta-analysis included 101 pre-1980 studies relating SES to achievement; Sirin's (2005) replicated and extended White's work using 74 post-1980 studies. Theoretically, SES is usually understood as a combination of parent income, occupational level, and educational level, but White found over 70 different variables in the operationalizations of SES he studied, including measures of home atmosphere, school resources, and other miscellaneous items (1982). Consistent with the Coleman Report, White noted that measures of home educational resources created a more predictive composite than the more traditional SES measures. For this reason, many researchers have added home resource variables to their operationalizations of SES (Sirin, 2005). Another choice is to unpack SES. Income, education, occupation, and home resources may have differing relationships with outcomes such as achievement, and the notion that these constructs can all be appropriately viewed as indicators of a broader SES concept has not been firmly established (Bollen, Glanville, & Stecklov, 2001).

Researchers and educators frequently use free-lunch status alone as a proxy for SES or to represent the better defined concept of family income. Students from families with incomes at or below 130% of the poverty level are eligible for free meals; those with incomes between 130% and 185% of the poverty level are eligible for reduced-price meals (Sirin, 2005). No

Child Left Behind reporting typically identifies free and reduced-price lunch students as a separate category (Kim & Sunderman, 2005). White's meta-analysis (1982) showed that 19 studies using family income as the only predictor of achievement had a higher mean correlation with test scores (M = .32) than studies using parental education only (116 studies; M = .19) or parental occupation only (65 studies, M = .20). Sirin's later meta-analysis (2005) found the three types of components (income, education, occupation) to be almost equally related to test scores, with mean effect sizes of .29, .30, and .28, respectively. A somewhat stronger predictor of test scores was the most commonly available measure of family income – free or reduced-price lunch status – with an effect size of .33 in 10 studies.

A Coleman study of the influence of background variables on student achievement was performed with a verbal aptitude test as the outcome variable. This raises the question of the relative strengths of relationships of SES to mathematics and verbal test scores. One meta-analyst found the relationship of SES with mathematics test scores to be stronger; the other found the relationship of SES with verbal test scores to be stronger. But in both meta-analyses, SES was related quite strongly to both verbal and mathematics achievement.[15]

The Coleman Report was ahead of its time in its careful delineation of within-school and between-school effects. Neither White nor Sirin reported on any studies that successfully made this separation. Each noted, instead, that the majority of studies used the individual as the level of analysis, while a minority of studies used an aggregate (such as a school, neighborhood, or district) as the level of analysis. They reported that results from aggregates

15 White (1982) found a stronger relationship between verbal scores and SES (M = .31) than between mathematics scores and SES (M = .25); Sirin (2005) found a stronger relationship between mathematics scores and SES (M = .35) than between verbal scores and SES (M = .32).

showed much stronger relationships between SES and achievement than results from studies of individual students.[16] Sirin chose not to analyze the two kinds of studies together, dropping the aggregated studies from his meta-analysis. He recommended the use of two-level modeling techniques in future studies, but apparently found none for his meta-analysis. In this way, the Coleman Report remains decades ahead of its time. The few studies that have successfully separated between-school and within-school effects since the work of Coleman et al. are discussed in a later section.

A methodologically cutting-edge analysis for the United Nations Educational, Scientific, and Cultural Organization (UNESCO) demonstrated that there is probably a significant relationship between reading performance and socioeconomic status in every country in the world (Willms, 2006). Willms calls these relationships "socioeconomic gradients" (p. 7). He analyzed data from the Programme for International Student Assessment (PISA), which is a collaborative initiative of member countries of the Organisation for Economic Co-operation and Development designed to assess the knowledge and life skills of fifteen-year-old youths. The literacy tests focus on the ability of students to use the knowledge they have learned in school. In 2000, 28 OECD countries participated in tests of reading skills; in 2002, another 14 non-OECD countries participated.[17] Willms demonstrated that all 42 of the participating countries had positive and significant relationships between student SES and reading test scores. Willms performed many other analyses, which are discussed in the appropriate sections that follow.

---

[16] The earlier sample (White, 1982) shows a mean correlation coefficient of .25 for studies with students as the unit of analysis and of .68 for those with aggregates as the unit of analysis. The later analysis (Sirin, 2005) yields very similar effect sizes of .28 and .67, respectively.

[17] The focus of the 2003 assessment was mathematics, but Willms analyzed only the reading results in his 2006 Working Paper.

Student Ethnicity and Test Scores

Coleman and his colleagues gave students tests in five subject areas. With regard to achievement, they found that "the order of the racial and ethnic groups is nearly the same on all tests. Following the Whites in order are Orientals, American Indians, Mexican-Americans, Puerto Ricans, and Negroes….The differences between Whites and the other racial and ethnic groups (excluding Orientals) is great indeed" (Coleman et al., 1966, p. 219). In the 40 years since that report, there has been more shuffling of categories and category names than substantial change in the results. This study and literature review focuses on the five broad ethnic categories of students identified by NAEP: Black, White, American Indian, Hispanic, and Asian, beginning with a discussion of the varied terms used to refer to these groups in the academic literature.

The Coleman Report (1966) speaks of Negroes; Secada (1992, p. 625) pointed out that this group of Americans has also been referred to by social scientists as Black, Afro-American, and African-American. For clarity, the language of the 2003 NAEP results is used here: this group of students is referred to as Black. Denotations will be changed as needed to mirror the language of the researchers in the literature review. The 2003 NAEP survey uses the term Hispanic. Secada (1992, p. 625) notes that Latino is another accepted term for this ethnic group, adding that it is often preferable to be more specific, as Coleman was. Coleman's focus on Mexican-Americans and Puerto Ricans is no longer quite as useful in the U.S. context because of increased immigration in the last four decades from other parts of Latin America. Following NAEP practice, this study will use the term "Hispanic," varying

only to match the language of literature review authors[18] or to be more specific to country of origin. NAEP includes a category for Asian / Pacific Islander. This study uses the term Asian for brevity, varying the language at times to match a source or to be more country-specific. The terms Native American and American Indian have been used interchangeably (Secada, 1992, p. 625). The latter designation is sometimes shortened to Indian. NAEP uses a category named "American Indian / Alaskan Native." This study uses the term American Indian as shorthand to refer to both of these ethnic groups, sometimes switching to the language of a source in this literature review or speaking more specifically about a particular group. Students whose ancestors emigrated to the U.S. from Europe are sometimes referred to as European American, sometimes as Caucasian, and sometimes as White. This study will generally follow the NAEP designation and refer to this group as White.

Since 1966, reviews of the data have consistently confirmed the findings of the Coleman Report: White and Asian students score much higher on standardized tests than Hispanics, American Indians, and Blacks. Ogbu (1978; 1992) has described the lower-scoring groups as involuntary and semi-voluntary minorities. His reports of cross-cultural research indicate that, worldwide, involuntary minorities tend to perform poorly in school.

Whites are the majority group in the U.S., at least for the next few years. The large majority of their ancestors immigrated to this continent by choice. The same is true of the large majority of Asians. Their immigration is typically more recent, but still almost completely voluntary. Blacks and American Indians have a different history; they are involuntary minorities. Blacks came to the U.S. as slaves. American Indians are living in a

[18] For example, the use by Coleman et al. (1966) of the term Oriental is outdated (American Psychological Association, 2001), but is used here in discussions of the Coleman Report.

nation built on land that was stolen from their people. Hispanics are an in-between grouping. Mestizo Hispanics share with American Indians a history of White theft of their land and property in both the United States and Latin America. Nevertheless, a large proportion of Hispanics have immigrated to the U.S. in search of opportunity. The Hispanic population, then, is best categorized as "semi-voluntary" (Ogbu & Simons, 1994).

Secada's (1992) review of the literature on ethnic differences in mathematics test scores shows how the pattern described by Ogbu plays out in mathematics scores on national surveys in the United States. Nine analyses of six national surveys[19] confirmed that, on mathematics tests, White and Asian students outperformed Black, American Indian, and Hispanic students. Separation of the Hispanic group into nationalities revealed important differences, but no Hispanic subgroup achieved scores near those of Whites and Asians, nor as low as Blacks. Tate's (1997) review of a similar collection of literature agreed that Whites outperformed Black and Hispanic students on both basic-skills and higher-order mathematics assessments. The small samples of post-Coleman surveys have limited the ability of researchers to address the performance of Asian and American Indian subgroups of students.

Since at least 1973, basic mathematics skills have improved for all ethnic groups. Overall, the improvements have been greatest for Blacks, American Indians, and Hispanics. Between 1973 and 1988, on NAEP's long-term trend assessment, which focuses on basic mathematics skills, the improvement was most rapid for these ethnic groups, leading to dramatic closing of ethnic test-score gaps. Since 1988, White and Asian improvement have matched or outpaced the improvement of Black, Hispanic, and American Indian students. Despite efforts

[19] Secada reviewed nine studies; data for the studies came from the 1973, 1978, 1982, and 1986 NAEP, the 1972 National Longitudinal Study, and the 1980 High School and Beyond Study.

to find socioeconomic, cultural, or educational patterns that match and explain this pattern, the reasons for the halt in progress toward equity remain complex and unclear (J. Lee, 2002).

Confirming evidence for this pattern is provided by a study that equated three other tests of basic mathematics skills given on national surveys in 1980, 1988, and 2002[20] using IRT[21] techniques (Cahalan, Ingels, Burns, Planty, & Daniel, 2006). The researchers showed an overall improvement of .0.40 standard deviations for high school sophomores in the 22-year period, with the largest increases shown by Black (.60 standard deviations), American Indian (.56), and Hispanic (.53) sophomores, but a different pattern for the 1990 – 2002 sub period in which the greatest improvements were made by American Indian (0.51) and White sophomores (0.21).

Recent grade 8 results from the Mathematics Main NAEP, a more balanced assessment of basic and higher-order skills than most assessments, show improvement by Asian, White, Hispanic, and Black eighth graders over the last decade and a half (see Table 1). However, ethnically-based gaps are not narrowing. A 33-point Black-White gap in 1990 became a 34-point gap in 2005. During the same period, the Hispanic-White gap increased from 24 to 27 points (National Center for Education Statistics, 2003, p. 13; Perie, Grigg, & Dion, 2005). Assuming that 10 to 13 NAEP points represents one grade level,[22] Blacks and Hispanics have, on average, scored two or more grade levels below Whites. It is important to

---

[20] The surveys, all conducted for the National Center for Education Statistics, were the 1980 High School and Beyond (HSB) survey, the National Educational Longitudinal Survey (NELS:88), and the Education Longitudinal Study (ELS:2002).

[21] IRT, or Item Response Theory, is a statistical methodology that puts test-takers and test items on the same scale. This allows valid judgements of item difficulty and comparison of scores across tests.

[22] Over the 1990 – 2005 period, the gap between the average eighth grade and the average fourth-grade score ranged from 41 to 50 points (Perie et al., 2005), leading to an estimate that the average score gain in one year is between 10.25 and 12.50 points.

remember, however, that these raw gaps include economic components along with the ethnic components because the data reported does not control for economic level.

Relationships between Economic and Ethnic Test Score Gaps

Both economic and ethnic test score gaps exist. This subsection examines the relationships that exist between these two kinds of test score gaps. The conclusion is that the two kinds of gaps are strongly related, but that neither subsumes the other. The two kinds of gaps are related because Blacks, Hispanics, and American Indians tend to have lower incomes than Whites, but close examination shows that any complete model of test scores must include both ethnicity and economic level.

Secada's (1992) review of the literature found that "poverty is more severely concentrated among African Americans and Hispanics than it is among Whites" (p. 633). Lubienski and Shelley (2003) provided recent detail using an SES variable created from 2000 NAEP data.[23] The upper SES quartile contained 32 percent of White students and only 8 percent of Black and Hispanic students. The upper half encompassed 61 percent of Whites, 25 percent of Blacks, and 20 percent of Hispanics. The lowest quartile held 12% of Whites, 48% of Blacks, and 52% of Hispanics.

[23] The SES variable incorporated types of reading material in students' homes, computer and internet access at home, extent to which studies are discussed at home, school lunch and Title I eligibility, and education level of mother and father.The authors did not disaggregate the SES variable to allow an investigation of its separate components.

Table 1. Grade 8 Main NAEP mathematics scale scores by group

	Assessment Year					
Demographic Category	1990	1992	1996	2000	2003	2005
National Mean	263	268	272	273	278	279
Ethnic Categories						
Asian/Pacific Islander		290		288	291	295
White	270	277	281	284	288	289
American Indian/Alaska Native					263	264
Hispanic	246	249	251	253	259	262
Black	237	237	240	244	252	255
Economic Level						
Ineligible for free/reduced-price lunch			277	276	285	288
Eligible for free/reduced-price lunch			250	253	259	262

Source: Perie et al. (2005)

In the 2005 NAEP survey, Whites comprised a much smaller share (35%) of the free- and reduced-price-lunch-eligible student population than of the ineligible population (77%). Conversely, Blacks, Hispanics, and American Indians comprised much larger shares of the free- and reduced-price-lunch-eligible population (29%, 30%, and 2%) than of the ineligible population (9%, 8%, and 1%). Asian students comprised a similar share of both populations (4% and 5%) (National Center for Education Statistics, 2004).

The strong relationships between ethnicity and economic level make it difficult to separate ethnic and economic effects on test score outcomes. This may be one reason that, in a survey of 3,011 mathematics education research articles, 112 considered race and 52 considered social class, but only 13 considered race and class together (S. T. Lubienski & Bowen, 2000). Nevertheless, ethnic identity is far from a perfect predictor of economic level. There are many poor children who are White; similarly, many Black and Hispanic students are not poor. These cases allow analyses that separate ethnic and economic factors. Secada (1992) described three statistically equivalent strategies that researchers employ to investigate both ethnicity and economic level: (a) considering income level as a variable within ethnic groups, (b) grouping simultaneously along lines of income level and ethnicity, and (c) considering

both income and ethnicity as predictor variables in a large-scale regression analysis. The different strategies can lead to different emphases in conclusions.

As reported earlier in this chapter, Coleman et al. (1966) applied the first strategy. Examining family background as a predictor variable within ethnic groupings, they concluded that family background was related to test scores within each ethnic group and that the economic level of the parents provided a small piece of that relationship. The second approach, grouping along lines of income and race simultaneously, is illustrated in Table 2, which is based on data collected for a profile of U.S. high school seniors from the 1992 National Educational Longitudinal Survey (NELS). NELS includes an index of SES. The data within any row in the left half of the table indicate that within a given ethnic group, students with lower SES are more likely to score in the lowest levels of mathematics proficiency. For example, among Asians, 26% of low-SES, 15% of mid-SES, and only 8% of high-SES seniors scored in the lowest proficiency levels. The same relationship between test scores and SES holds for each ethnic group. The right half of the table shows a related pattern: within any given ethnic group, higher SES students are more likely than lower-SES students to score at the highest proficiency levels. At the same time, the data within any column indicates that within each level of SES, the ordering of mathematics proficiency by ethnicity remains constant – Asians are least likely to score at the low levels and most likely to score at the high levels, followed by White, Hispanic, and Black students. The clarity of these patterns shows that any explanation of mathematics test scores in the United States must take account of both SES and ethnicity. Neither demographic category can explain these patterns by itself.

Table 2. Mathematics proficiency crosstabulation by ethnicity and socioeconomic status

Ethnicity	Percent of students in each ethnic/socioeconomic category scoring at lowest levels of mathematics proficiency			Percent of students in each ethnic/socioeconomic category scoring at highest levels of mathematics proficiency		
	Low SES	Mid SES	High SES	Low SES	Mid SES	High SES
Asian	26	15	8	23	41	65
White	40	22	8	18	35	59
Hispanic	51	35	17	13	25	44
Black	60	45	26	5	16	27

Note. Adapted from (Tate, 1997). Original data from (Green, Dugoni, Ingels, & Camburn, 1995). Based on 1992 NELS:88 second follow-up survey of high-school seniors. The NELS exam has levels of proficiency from below basic to level 5. The lowest levels of proficiency are below basic and level 1; the highest levels of proficiency are 4 and 5.

In a series of analyses of NAEP mathematics test scores presented at the American Educational Research Association, Lubienski (2001; 2002; 2003) mined data from similar cross-tabulations. Between 1990 and 1996, she found a grade 8 mean mathematics test score gap between White and Black students that was large,[24] widening, and present at both highest and lowest SES quartiles. Both economic and ethnic differences were clearly documented by her tabulations (S. T. Lubienski, 2001). Her look at 2000 data with free/reduced-price lunch eligibility tabled against White, Hispanic, and Black ethnic categories (see Table 3) showed the same clear pattern of test-score advantages based on ethnicity within economic category along with test-score advantages based on economic level within ethnic group (S. T. Lubienski, 2002).

[24] Lubienski suggested that a 10-point NAEP achievement gap is roughly equivalent to one grade level, and therefore that the 39 point Black-White grade 8 NAEP mathematics test score gap she found in 1996 was the equivalent of nearly four grade levels. Furthermore, the average score of Black 12th graders was lower than that of White 8th graders.

Table 3. Mean grade 8 NAEP mathematics scores by ethnicity and lunch eligibility, 2000

Ethnic Category	Eligible for free/reduced-price lunch	Not eligible for free/reduced-price lunch	Economic test score gap
White	270	289	19
Hispanic	246	263	17
Black	242	255	13
White-Hispanic Gap	24	26	
White-Black Gap	28	34	
Hispanic-Black Gap	4	8	

Note. Adapted from Lubienski (2002).

Limiting themselves to Black and White, Phillips et al. (1998) used the third approach to simultaneously investigate the effects of ethnicity and economic level on test scores – a regression with ethnicity and economic level as predictor variables. Their outcome variable was performance on the Peabody Picture Vocabulary Test scores (PPVT) for 1,626 children from the Children of the National Longitudinal Study of Youth (CNLSY) survey. A regression including race as the only predictor explained 22% of the test score variance. A variety of codings of family income each added up to 5% to the variance explained. The addition of parents' accumulated wealth added no explanatory power when income was included. Parental occupational status, household size, number of parents, and mother's work, taken together, were more predictive than family income, bringing the variance explained up to 35%. Re-addition of the family income and wealth variables added nothing to the prediction. The addition of a rich set of mother's experiences brought the explanatory power up to about 50%; measures of the mother's cognitive skills brought it up to 67%. As other variables were added, the importance of race declined. A 16-point deficit for Blacks in the race-only model became a 14-point deficit in the various income-including models and an 8-point deficit with the broader SES variable. In the most complete model, this 8-point ethnic deficit was much larger than a 5-point difference between families with average income below $12,500 and those with average income over $50,000, but both remained significant.

All the studies cited above agree that although family income (or SES) and ethnicity are strongly related to each other, neither of these factors alone can predict all of the observed differences in test scores. Economic level and ethnicity must be considered as separate, though strongly related, concepts.

Section 2. Within-school, Between-school, Total, and Composition Effects

The most sophisticated post-Coleman analysis of SES and ethnic test score gaps with NAEP data was provided in *Examining Instruction, Achievement, and Equity with NAEP Mathematics Data* (S. T. Lubienski, 2006). The results are shown in Table 4. This study validated the findings from the previous section: ethnicity and economic level (or SES) are each independently important variables in the prediction of student test scores. It improved on those findings by using a two-level regression to explore the levels at which these effects occurred. The basic equation for this kind of separation is: within-school effects + composition effects = between-school effects[25].

Lubienski's two-level regression incorporated the SES and ethnicity of individual students as within-school predictors and school socioeconomic and ethnic compositions as between-school predictors. The outcome variable was grade 4 NAEP 2000 mathematics test score. Within-school controls were student gender and disability status. Private school status was a between-school control.[26]

---

[25] Willms (2006, pp. 49-51) provides two key formulas. First, he states that between-school effects are the sum of within-school effects and composition effects. Second, he cites Alwin (1976) and shows that the "overall gradient slope," which is here referred to as a "total effect," is equal to η^2 (Between-school slope) + (1 - η^2)

[26] The value 1 for "Black" or "Hispanic" indicated membership in the named ethnic group. The comparison group, occupied by all students with values of 0 for both Black and Hispanic, contained, by default, Asian, White, American Indian, and all other categories of students. "School Race/Ethnicity" was based on the percentage of White/Asian students in the sample; the variable was logarithmically transformed to create a standardized variable (normal distribution with mean = 0 and standard deviation = 1). Factor analysis was used to create a comprehensive, standardized SES variable. This variable incorporated types of reading material in

As shown in Table 4, within an average school, when all other variables are controlled, Black students are expected to score 17.1 points below non-Black/non-Hispanic students. Similarly, Hispanic students are expected to score 12.5 points below non-Black and non-Hispanic students. These are *within-school effects*; they are a weighted average of the effects found within each of the schools in the sample. These effects are related to characteristics of the students, their families, and the ways that these characteristics interact within an average school. They are not related to any differences in the characteristics of the schools attended by these different ethnic groups.

Table 4. A two-level model of NAEP 2000 grade 4 mathematics test scores

Variable	Parameter Estimate
Level 1 (within schools)	
Intercept	236.6***
Black	-17.1***
Hispanic	-12.5***
Student SES	7.6***
Boy	3.8***
Disability	-30.3***
Level 2 (between schools)	
School SES	6.1***
School Race/Ethnicity	not significant
Private School	-4.8***
Variance Components	
Variance between schools	83.6
Variance within schools	454.2
Intraclass Correlation Coefficient (ICC)	.16

Note. Adapted from Lubienski (2006). N=9999 students and 611 schools.
*p<.05; **p<.01; ***p<.001

students' homes, computer and Internet access at home, extent to which studies are discussed at home, and eligibility for school lunch and Title 1. Another standardized composite variable was created for school SES. The measures of this latent variable were a school aggregate of the individual SES variable and the percentages of students eligible for free/reduced-price lunch and Title 1. Binary variables for gender and disability status were also used at the individual level. School sector (public or private) was a school-level covariate.

In this well-controlled model, the ethnic composition of the school[27] showed no significant independent impact on test scores. For this reason, the between-school ethnic effect may be considered to be equal to the within-school ethnic effect. Controlling for all other factors, an average all-Black school is expected to score 17.1 points below an average school with no Black or Hispanic students (the between-school effect), just as an average Black student is expected to score 17.1 points below an average White student within any given school (the within-school effect). The *total effect*, which would be estimated by a standard regression of NAEP grade 4 test score on ethnicity with students as the unit of analysis, is an average of the within- and between-school effects, weighted for the degree of ethnic segregation in the system (Willms, 2006).[28] In this case, the total effect was identical to the within-school and between-school effects: 17.1 points. With all other controls in place, the degree of ethnicity-based segregation in the nation's school system was not shown to be relevant to grade 4 mathematics test scores.

While Lubienski's results did not demonstrate a composition effect for ethnicity, they did show a socioeconomic composition effect. Within an average school and controlling for all other variables, a student with an SES one standard deviation higher than another student would be expected to have a NAEP score 7.6 points higher. This is a 7.6-point within-school SES effect. The parameter estimate for School SES was 6.1. Because School SES was centered on the national mean (grand-mean centered, in statistical language), this is an estimate of the composition effect of SES. If school A has a student-body average SES 1

---

[27] Lubienski's "school race" variable is engineered to have a mean of 0 across the sample (S. T. Lubienski, 2006). It is therefore grand mean centered, and its coefficient can be interpreted as a composition effect (Raudenbush & Bryk, 2002).

[28] The effects estimated by the White (1982) and Sirin (2005) meta-analyses are almost all total effects because they focus on the student level with no two-level separation.

standard deviation higher that of school B, then a student in school A is expected to score 6.1 points higher on the grade 8 NAEP mathematics assessment than an identical student in school B[29] (Raudenbush & Bryk, 2002).

The *between-school effect* of SES is the sum of these two effects; in this case, 13.7 points. According to Lubienski's study and controlling for all other factors, a school that is 1 standard deviation above the national average in School SES would be expected to score 13.7 points higher than an average school on the grade 4 NAEP mathematics test. The within-school effect, 7.6 of these points, would be due to the SES level of the students seen as individuals. The other 6.1 points, the composition effect, would be due to the beneficial effects that being in a school with a high socioeconomic level has on all of the students in the school.

Possible reasons for composition effects will be discussed in more detail later in this chapter. Lubienski did not estimate total effects, but because of the mathematical relationships of the effects, they must lie somewhere between the within-school effect (7.6) and the between-school effect (13.7), depending on the degree of socioeconomic segregation in the system. In a completely segregated system, the total effect matches the between-school effect. In a fully integrated system, the total effect would be much smaller, matching the within-school effect because there is no variance in the composition of schools.

[29] In reality, it is very possible that this SES composition effect might differ for different ethnic groups (Secada, 1992). Lubienski's model, and the model proposed by this study, assumes that any such differences are insignificant.

Economic Composition Effects

The remainder of this section describes some of the other studies that distinguish within-school effects, composition effects, between-school effects, and total effects. In particular, it focuses on the question of economic composition effects: Does the socioeconomic composition of a school predict that school's effectiveness at generating high test scores for all groups of students in the school?

Myers (1985) used path analysis on national datasets and demonstrated an SES composition effect. He found that concentration of poverty within schools is related to mathematics scores, even controlling for family SES, student-level gender, race/ethnicity, family structure, maternal work, and number of siblings. By controlling for family SES in a study of school-level SES, Myers distilled a pure composition effect.

Using a large dataset from the Educational Testing Service (ETS), Everson and Millsap (2004) estimated a multi-level structural equation model in which family socioeconomic background (parental income and education) predicted high school achievement and participation in extracurricular activities, and all three variables predicted SAT scores within schools. The researchers aggregated these three variables to the school level and added school ethnic and economic composition,[30] size, and locale as control variables to create a between-schools model.

At the between-schools level, a 1-standard-deviation increase in family socioeconomic background predicted a 60.1-point increase in SAT mathematics score, a sizeable

[30] The inclusion in the between-school model of both aggregated parental income (as part of the aggregated family background latent variable) and percentage of students with free or reduced-price lunch status appears to be a misspecification, with the effect of reducing the predictive power of both. It may be for this reason that no significant effect was found for the percentage of free/reduced-price lunch students in the school.

composition effect. A significant estimate for the between-schools ethnic composition variable was also found: a 10 percentage point increase in the number of minority students was associated with a 7.6 point drop in SAT mathematics score. Unfortunately, the model did not include a student-level, within-school estimate of ethnic effect, so the between-school coefficient is almost surely an overestimate of the actual ethnic compositional effect. In fact, it is an estimate of the overall between-schools effect, including the effect of student-level ethnicity as well as the effect of school-level ethnic composition.

Willms's (2006) international multi-level models showed significant SES composition effects on PISA test scores in every country tested, with considerable variance between countries in the intensity of the effects. "Those countries with the best results – high and equitable student performance – with very few exceptions have low levels of between-school segregation" (Willms, 2006, p. 50). To avoid the political challenge of reducing segregation, a nation might instead attempt to lessen the magnitude of the composition effect by "bolstering the achievement levels of low-SES schools," but "this is difficult to achieve because when low-SES or low-ability students are concentrated in particular schools, it is difficult to maintain high expectations, establish a positive disciplinary climate, and attract and retain talented teachers" (p. 51).

Understanding the importance of SES composition, some school districts have economically desegregated their schools, hoping to reduce between-school differences that widen socioeconomic test score gaps. One such district is Wake County, North Carolina (Flinspach & Banks, 2005). A study of Wake County and other large North Carolina school districts found that within each district, the economic composition of a school is predictive of the passing percentage for free/reduced-price lunch students and also for students ineligible

for free/reduced-price lunch. But overall, Wake's socioeconomic integration has led to dramatically smaller between-school passing percentage gaps than are found in the state's four other largest districts (Regan, 2005). The county's positive results are not surprising, given the preponderance of evidence showing that a school's socioeconomic composition predicts its effectiveness.

Ethnic Composition Effects

One of the key findings of the Coleman Report was that the achievement of minority students was higher in racially integrated schools. The presumption was that school ethnic composition had an effect on the learning of the students in the schools and that separate schools were inherently unequal. This and similar findings led to a major social experiment in the United States in the late 1960s and early 1970s, the racial desegregation of Southern schools. In 1954, the Supreme Court ruled that "separate but equal" schools for Black and White children were unconstitutional. Fifty years later, the ruling was commemorated but not celebrated by Black scholars or policymakers (Street, 2005). The somber tone was because the ideals of the ruling, that the nation's schools would become ethnically integrated and equally effective, were never realized. The ruling succeeded in ending legally enforced school segregation within districts and brought massive desegregation to the nation's schools, especially in the South. In 1954, 40% of the nation's Black students and virtually 100% of those in the South attended segregated schools. Now, the majority of Black students attend integrated schools (Street, 2005). But the nation's schools have never been fully desegregated. By the late 1980s, they had begun to resegregate due to a combination of judicial regression, housing segregation, segregation academies, and white flight (Associated Press, 1999; Boger & Orfield, 2005; Ladson-Billings, 2004; Orfield, 1996, 2001; Street,

2005; Wright, 2006). America's Black, Hispanic, American Indian, and White students are still likely to attend schools that are far from ethnically and economically balanced (Orfield, 1996).

Desegregation offers a natural experiment for assessing the effects of changing the ethnic composition of schools. Literature about ethnic desegregation since the landmark 1954 *Brown vs. Board of Education* Supreme Court decision in the U.S. is reviewed here, with a particular focus on the basic ethnic composition effect question: "Does the ethnic composition of a school have an effect on student achievement in that school, above and beyond the effects of the ethnicities of the individual students in the school?"

De facto ethnic segregation exists and has always existed in the United States (Spring, 2004). But for a brief period (1968 – 1972), in a distinct region (the South), and for a distinct pair of ethnic groups, the level of segregation declined sharply. The share of Southern Blacks attending schools that were more than 90 percent minority plummeted from more than three-quarters to about one-quarter over this four year period. Between 1968 and 1991, Black-White segregation also decreased (though much more gradually) in the West and Midwest, while actually increasing in the Northeast, which moved from the least to the most segregated region during those years (Grissmer, Flanagan, & Williamson, 1998). Resegregation began in the late 1980s, most strongly in the South. It is perhaps not coincidental that the test score gains for Black adolescent students in cohorts entering school between 1968 and 1980 were about 0.6 standard deviations for reading and mathematics combined. The gains were largest in the Southeast and smallest in the Northeast.[31] According

[31] Averaged across subjects and ages, Black NAEP scores increased by about 0.80 standard deviations between the 1970s and 1992 in the South, by 0.65 standard deviations in the West, 0.55 standard deviations in the Midwest, and 0.35 standard deviations in the Northeast (Grissmer et al., 1998, Figure 6.6, p. 193).

to the analysts who presented these results, "such large gains over such a short period are rare; indeed, they may be unprecedented" (Orfield, 1996, p. 194). The results are attributed in part to desegregation.

Black-White desegregation is the one historical variable that tracks most closely with long-term trends in achievement gaps (J. Lee, 2002). The Hispanic-White gap never narrowed as dramatically, with progress ending in about 1982. Perhaps this is related to Hispanic-White segregation never having a dramatic decline (J. Lee, 2002; Orfield & Yun, 1999). These historic trends suggest positive effects for ethnic desegregation on Black students' test scores. More evidence comes from research focusing on specific desegregation efforts.

A 1978 review of the quantitative literature on the results of integration efforts (Crain & Mahard) looked at 39 studies of desegregation plans involving mandatory reassignment of Black students. Twenty-four reported gains in test scores; five reported losses. The average change was about one-half of a grade level equivalent. More methodologically advanced studies showed larger improvements. The scores usually improved in the short run, and always in the long run.

A more recent study, using SAT scores of Black students across different metropolitan areas with detailed controls for family background of individual test-takers; school-level controls for selective test participation; and city-level controls for racial composition, income, and region found "robust evidence that the black-white test score gap is higher in more segregated cities" (Card & Rothstein, 2006, abstract), but suggested that the cause is neighborhood segregation more than school segregation.

The Gautreux program in Chicago (Rosenbaum, 1995) provided experimental evidence that neighborhood and school have a significant impact on student outcomes. Low-income Chicago families were randomly assigned to housing in predominantly minority inner-city or predominantly white suburban Chicago. With similar grades, the children in the families assigned to suburban housing were less likely to drop out, more likely to be in a college track, more likely to go to college, and more likely to go to four-year college.

Researchers generally agree that school ethnic composition is strongly related to school economic composition (Boger & Orfield, 2005; Card & Rothstein, 2006; Cashin, 2004; Darling-Hammond, 2006; Kozol, 2005; J. Lee, 2002; Oakes, Rogers, Silver, & Goode, 2004; Street, 2005; Teranishi, Allen, & Solórzano, 2004) and that this provides at least some of the explanation for the positive effects of ethnic integration on the achievement of involuntary minority students. For example, one careful analysis of national data finds that

> When African-American and Latino students are segregated into schools where the majority of students are non-white, they are likely to find themselves in schools where poverty is concentrated. This is of course not the case with segregated white students, whose majority-white schools almost always enroll high proportions of students from the middle class. This is a crucial difference, because concentrated poverty is linked to lower educational achievement….When school districts return to neighborhood schools, white students tend to sit next to middle-class students but black and Latino students are likely to be next to impoverished students. (Orfield & Yun, 1999, p. 3)

Because of their tight links, one cannot assess the ethnic composition effect without controlling for the economic composition effect. Excepting Lubienski (2006), only two studies have effectively controlled for economic segregation. A group of Florida researchers (K. M. Borman et al., 2004) analyzed comprehensive data collected by the Florida Department of Education. Using ordinary least squares regression with a strong set of control

variables,[32] they reported that both ethnic and economic composition were predictive of passing percentages on state tests in reading and mathematics at elementary, middle, and high school levels. The effects they report, however, are between-school effects because they are comparing school-level passing percentages. As such, they confound the within-school and composition effects, leaving this otherwise strong study unable to confirm or deny the existence of ethnic (or economic) composition effects and therefore unable to truly support their inference that the Florida schools should integrate more fully.

Along with the Lubienski (2006) results already described, the best evidence on ethnic composition effects comes from an analysis of Louisiana high school sophomore test scores (Bankston & Caldas, 1996). The study provides a strong set of controls at two levels and concludes that "the degree of minority concentration has a powerful negative influence on achievement test results, that this influence does not appear to be explained by socioeconomic factors or other factors, and that both whites and African Americans are negatively affected by degree of minority concentration" (Bankston III & Caldas, 1996, p. 1).

The dataset, provided by the Louisiana Department of Education, included all of the 40,041 White and Black non-special education public school sophomores who provided useable test score data in 1990. Rich demographic information was available from these students. The outcome variable was a composite of the three Louisiana state tests taken by the group.[33] The researchers included a wide variety of both individual[34] and school-level variables[35] as predictors in their regression analysis.

---

[32] The control variables were per-pupil expenditures, instructional quality, percent mobility, percent Hispanic, average class size, and percent receiving free or reduced-price lunch.

[33] Louisiana sophomores are tested in mathematics, English Language Arts, and written composition as a part of the state graduation requirement. A composite of the three tests was created with principal components analysis. It was determined that the three scores in fact measured a single underlying construct that the

The study began with a series of student-level analyses. Gender (.072) and LEP status (.013) were both significantly related to achievement, but race (-.358) had a far stronger effect in a model containing only those three variables. Adding the five indicators of student time use produced interesting results, but did nothing to explain away the individual effect of race. In fact, the race coefficient strengthened to -.377. The addition of free lunch status (with a significant -.120 coefficient) to the model attenuated the effect of individual race only slightly, to -.313. Finally, the researchers added parental SES to the model. It became the second-strongest indicator, but had no effect on the coefficient for student race. This may be because it was somewhat collinear with the free lunch status variable already in the model. The coefficient for free lunch status remained significant, but was reduced greatly, to -0.69.

The subsequent series of models added school-level variables in steps. The first school-level variable added to the model was percent minority in the school. This variable decreased the strength of the individual race variable (to -.255) by explaining a good share of that variable's power itself. The addition of the five student time use variables (aggregated to the school level) to the model had little effect. Three of the new variables were significant, and the strength of the percent minority variable increased. When the percentage of free lunch students and mean parental SES were added to the model, the importance of school ethnic composition only increased, despite the fact that both new variables were significant.

---

researchers called Academic Achievement. This single factor accounted for about 73 percent of the overall variance in the analysis.

[34] The individual variables were: (a) binary indicators of race, sex, free/reduced-price lunch status, and limited English proficiency, (b) five variables indicating student time use: daily hours spent watching television, reading, and doing homework, and weekly hours spent working and in organized activities, and (c) a composite of parents' educational and occupational level used to measure parent SES.

[35] The school-level variables were created by taking the school-level mean of each individual-level variable noted above.

A four-model sequence was then used with Black students and again with White students in order to see if the factors predicting test scores differ dramatically between the two groups. There are few differences. In particular, school economic level and ethnic composition are significantly predictive of test scores in the expected way for both groups of students, even with the full set of controls at both school and individual level. Both Black and White students are affected negatively by attendance in a school with a high proportion of Black students; the effect is larger for Black students.

Although it uses variables at both student and school levels, this is not a true two-level model because the variance is never separated into within-school and between-school components. All variables at both levels are viewed as predictors of student-level variance. This methodological flaw leads to underestimation of standard errors and a concomitant risk of increased Type I errors.[36] Inefficient and biased estimates of the compositional effects of interest may also result (Raudenbush, 2002, p. 102). Nevertheless, the inclusion of the aggregated school-level variables is a major strength of this model compared to most of the quantitative literature in the field.[37]

The Louisiana study suggests that ethnic composition effects exist; the Lubienski (2006) NAEP study suggests that they do not. The difference may be a Type I error in the Louisiana

---

[36] A Type I error is committed when a researcher falsely rejects the null hypothesis of no relationship between variables. Most of the relationships found in this study were declared significant at the .01 level, suggesting that one such relationship in a hundred will be spurious on probabilistic grounds. The underestimation of standard errors created by the lack of a true multilevel model increases this error rate by an unknowable amount. The strength of the relationships reported makes this problem less worrisome than it might otherwise be. It is also important to note that this problem is not introduced by the researchers' attempt to use aggregated school-level variables; it is a feature of any person-level analysis of students within schools data.

[37] This study could also be strengthened with structural equation modeling. It is a complicated model including many presumed causal relationships. SEM would allow these paths to be modeled and tested.

study, introduced because of the failure to perform true multi-level modeling. Alternatively, it may be that ethnic composition effects exist in Louisiana, but not in the nation as a whole, or that ethnic composition effects arise sometime between fourth-grade and eighth grade. In any case, it is worthwhile to investigate the existence of and possible reasons for ethnic composition effects.

Section 3. Possible Reasons for Economic and Ethnic Composition Effects

Previous sections suggest that schools with more involuntary minority and poverty students may be, on average, less effective. This section focuses on five possible reasons. An effort is made to differentiate the effects of poverty from the effects of membership in various ethnic groups, insofar as such differentiation is facilitated by the literature. The first four reasons are briefly discussed below; the final reason, differential Full-School Engagement, is the primary subject of the remainder of this literature review.

Student Needs

"Establishing an optometric clinic in a school to improve the vision of low-income children would probably have a bigger impact on their test scores than spending the same money on instructional improvement" (Rothstein, 2004, pp. 9-10). Rothstein's example makes the point that low-income students come to school with many needs that make the job of teaching them particularly challenging. His bottom line is that even if they were fairly distributed, good teachers and other school resources would not be enough to generate equal results (p. 3). The resources of poverty schools may be stretched thin by the needs of the students attending them.

Using data from a variety of sources, Rothstein documents social class differences in childrearing, vision, hearing, oral health, lead exposure, asthma, medical care, parental

alcohol use and smoking, incidence of low birth weight, nutrition, housing, and student mobility. According to Rothstein, these are some of the reasons that, in 40 years of post-Coleman research, "no analyst has been able to attribute less than two-thirds of the variation in achievement among schools to the family characteristics of their students" (p. 14).

A summary of literature reviews on the correlates of achievement for the Educational Testing Service (P. E. Barton, 2003) found 14 correlates of achievement test scores. All of the correlates were also correlated with student ethnicity; almost all were correlated with student economic level. Seven of them are factors over which schools have little control, i.e., birth weight, lead and other environmental hazards, hunger and nutrition, parent availability, amount the child was read to at an early age, amount of television watching, and student mobility.

In 2007, *Education Week's* annual review of the quality of the nation's schools underwent a major revision – the prior focus on K-12 schooling was replaced with a focus clarified in the title of the special issue: "From Cradle to Career" (Olson, 2007). The point is that educational preparation for a competitive world starts with birth and continues at least through college. The chapter on pre-school child well-being points to the educational significance of large inter-state disparities in birth weight, family income, parental education, parental employment, linguistic integration, child health insurance, and spending on child-care services. This body of research may be best summarized in the words of Gary Orfield (quoted in Street, 2005, p. 107): "Where students come to class hungry, exhausted, or afraid, when they bounce from school to school as their families face eviction, where they have no one at home to wake them up for the bus, much less look over their homework, not even the

best-equipped facilities, the strongest curricula, and the best-paid teacher can ensure success."

The implication is that schools serving large populations of low-income and involuntary minority students must address a wide variety of needs that schools serving more advantaged students do not face. It is reasonable to believe that these needs draw attention from the academic needs of the students, individually and as a whole. These challenges may be a part of the reason for economic and possible ethnic composition effects on test scores.

Varieties of School Resources

Teachers in schools serving lower income or involuntary minority students are more likely to report insufficient resources to do their jobs (Strutchens et al., 2004). This may be in part because their jobs are complicated by the factors mentioned above, but there is substantial evidence that it is also because their schools actually have less resources. The following sections will provide evidence that schools serving less low-income and involuntary minority students tend to have (a) more money per student, (b) better facilities, (c) more and better instructional materials, (d) stronger curriculum offerings, (e) more emphasis on reasoning, (f) better teachers, and (g) a more favorable school disciplinary climate. Furthermore, the quantity of each of these resources is predictive of average school achievement, at least in some degree. Taken together, evidence for the importance of these resources and for their maldistribution implies that the resources mediate the effects of school ethnic and economic composition on test scores. Many resource inequality studies combine this broad range of categories; these studies are discussed first, with a focus on the question: is unequal resource distribution part of the reason for ethnic and economic composition effects? Second, studies more specifically focused on money, facilities, materials, curriculum offerings, and reasoning

are presented in subsections. A separate section is reserved for consideration of the resource of teacher quality. A final resource, labeled school climate, is discussed later in this literature review because of its close relationship with Full School Engagement, the central construct of the study.

Do Resources Matter?

The two-level analyses of the Coleman Report concluded that overall school quality could account for no more than a quarter of the variation in student test scores,[38] depending upon the ethnic/geographic group in question. Furthermore, controlling for the economic and educational backgrounds of the school's students, measured characteristics of the schools attended[39] by the ninth grade sample accounted for less than 8%[40] of the variance in test scores for every ethnic/geographic group (Coleman et al., 1966, p. 306, Table 3.23.2). The researchers concluded that "differences in school facilities and curriculum, which are the major variables by which attempts are made to improve schools, are so little related to differences in achievement levels of students that, with few exceptions, their effects fail to appear even in a survey of this magnitude" (p. 316).

One weakness of the Coleman Report is that it ignored the variation in school quality that occurred between ethnic/geographic groups because it looked separately at each category of

---

[38] At grade nine, they found that roughly 6% of the variance in student verbal scores was between schools for Oriental American students, 9% for Northern White and Southern White students, 13% for Northern Black students, 16% for Mexican-Americans, 20% for Southern Negro students, 21% for Puerto Rican students, and 24% for Indian American students (Coleman et al., 1966, p. 296).

[39] For the ninth grade sample, Coleman et al. measured the following school characteristics: per pupil expenditure on staff, volumes per student in the library, science lab facilities, extracurricular activities, presence of accelerated curriculum, comprehensiveness of curriculum, use of tracking, movement between tracks, guidance counselors, and school location (city, suburb, town, country).

[40] Measured characteristics of the schools attended accounted for about 8% of the variance in the test scores of Southern Negroes, less for all other categories of student.

student. Many subsequent education production function studies avoided this problem, sometimes adding a different problem by ignoring racial and ethnic categories altogether, but arrived at similar conclusions. An influential review (Hanushek, 1986) counted positive, negative, significant, and non-significant coefficients in 65 education production function studies, concluding that "there appears to be no strong or systematic relationship between school expenditures and student performances" (p. 1162). Two re-reviews of the same set of analyses used different statistical methods[41] and came to a different conclusion: "School resources are systematically related to student achievement and…these relations are large enough to be educationally important" (Hedges & Greenwald, 1996).

A re-analysis of the same set of studies (Dewey et al., 2000) suggested that claims for the ineffectiveness of school resources were based on a misspecification. Two-thirds of the studies had included income as a parental input in their simple regression.[42] The problem with this specification is that, through the schooling and housing markets, parental income is a major determinant of the school the child attends. In this manner, parental income is one of the determinants of school resources. If it is placed in a list of predictor variables alongside school resources in a simple regression model, it will lead to an underestimate of the effectiveness of those resources. The re-analysis showed that in studies that did not include parental income, significantly positive school input coefficients were 39% more common than in incorrectly specified studies. Their meta-analysis found clear support for the claim that teacher education, teacher experience, teacher salary, teachers per pupil, expenditures per

[41] The researchers used meta-analysis and a simple look at median coefficients instead of counting significance of results, which is as much affected by power of the study as it is by the strength of the effect. (Hedges & Greenwald, 1996)

[42] The Coleman Report also included student economic background as an input in a simple regression model.

pupil, and other teacher characteristics are all positively related to educational outcome. An alternative to leaving parental income out of the specification is to include it in a more complete structural model, modeling its role as predictor of resources as well as its role as direct predictor of student test scores (Levačić & Vignoles, 2002).

School Resources as Mediators

Willms (2006) describes an ambitious attempt to uncover the mediating role of school resources on school effectiveness, but falls prey to the specification error described above. By putting the resource variables into simple regression equations (albeit at two levels) instead of modeling them more accurately as mediators in an SEM framework, the coefficients for their effects are downwardly biased. Using PISA 2000-2002, Willms found that in the average country, student-teacher ratio, teacher educational level, morale of teachers and commitment, student-teacher relations, disciplinary climate, and student use of resources were significantly predictive of adjusted between-school test score averages.[43] It is possible that a more accurate modeling of the role of resources as a mediator between school composition and test score outcomes would have yielded even more resources that are significantly predictive of school effectiveness, as well as clarifying their mediating role.

Another approach to looking for mediators of economic and ethnic effects on mathematics test scores is to identify resources that are associated with higher test scores and noting the degree to which various groups have access to those resources. Using the 1992 NAEP Trial State Assessment and a well-controlled multi-level approach, Raudenbush, Fotiu, and

[43] The quality of school infrastructure, the number of computers per student, and the amount of professional development received by teachers were all found to be insignificantly related to test scores after other variables were controlled. No school policy variables were significant.

Cheong (1998) confirmed that disciplinary climate, the offering of algebra, teacher undergraduate mathematics major, and teacher emphasis on reasoning were strong predictors of mathematics test scores among all groups of students. They also found that these resources were more available to White students, Asian students, and the children of more highly educated parents. They did not use a structural equation model and therefore were unable to put the two parts of the mediation effect together, failing to fully estimate the mediating role played by these resource differences.

One rare education production function includes both structural equation and two-level modeling (Wenglinsky, 2002). It is also advantaged[44] by the strong set of classroom-level data available in the grade 8 NAEP mathematics database. It is because of these advantages, Wenglinsky believes, that he was able to find resource effects on mathematics test scores that are comparable in size to student background effects. SEM allowed him to simultaneously estimate the effects of school composition on resources, of resources on test scores, and of school composition on test scores. Because it shares key methodological elements with the current study, Wenglinsky's study is discussed at the end of this chapter.

Money

The most basic of resources for education may be money. The best evidence is that schools serving higher-income students have access to more dollars per student and that this makes a difference in the quality of the education received by the students in those schools. There is much less equality in per-pupil expenditures in the United States than in European

[44] Researchers on both sides of the "do resources matter" divide (Hanushek, 1996b; Hedges & Greenwald, 1996) note that the education production function literature could be improved by access to more classroom-level data. Another pair of researchers, using data that allowed students to be linked to particular teachers, were able to show a significant effect of teacher qualifications. They did not find, however, that the lack of these variables biased the results of other education production functions (Goldhaber & Brewer, 1997).

and Asian countries where spending is centralized and equal (Darling-Hammond, 2004). The differences are primarily because local property taxes are a primary source of dollars for schools in the U.S. (Education Trust, 2006; Evans, Murray, & Schwab, 1997; McCabe, 2006a). Districts with large tax bases are heavily advantaged. For example, in Texas, taxpayers in the 100 wealthiest districts paid 47 cents per $100 of property valuation, raising $7000 per student in the district. At the same time, taxpayers in the 100 poorest districts paid 70 cents per $100 and raised just $3,000 per student (Kozol, 1991, p. 225). Court decisions in many states over the past few decades have required some degree of equalization between districts. These decisions have led to increased state-level spending, aimed primarily at poorer districts. Nevertheless, the ratio between 5th and 95th percentile per-pupil spending in 1972 was 2.45. It was no different in 1992 (Evans et al., 1997). By 2004, the ratio had climbed to 2.72.[45] In 39 of 49 states, property-poor districts receive less from state and local revenues than wealthy districts (S. T. Lubienski, 2006). The Federal government adds 9% of the funding stream, but the formula for the Title I program designed to equalize spending is itself biased toward wealthier states (Education Trust, 2006).

These statistics make clear that school funding tends to favor economically advantaged students. An argument that these differences are an important part of the reason for (or, statistically, a mediator of) the relationship between school economic composition and test scores requires, however, evidence that more money tends to lead to higher scores. A recent review of the education production function literature (Jefferson, 2005) sees this question as still unanswered. Money can surely be used to help to improve achievement, but it is unclear

[45] "Five percent of regular districts had total revenues per pupil of $6,621 or less, while 5 percent had total revenues per pupil of $18,071 or more" (Sable & Hill, 2006, p. 5).

whether schools use it to effectively do so. In particular, it is unclear whether the differential funding documented in the previous paragraph is a major part of the explanation of economic test score gaps.

NAEP does not include spending variables. Therefore, it is rarely used to investigate these questions. One researcher, however, supplemented the NAEP database with census data and found that, controlling for a large list of other key factors, states that spend more on education tend to achieve higher mean test scores (Grissmer et al., 2000). This debate is not fully resolved, but it is at least reasonable to suppose that differential monetary resources provide part of the explanation for test score gaps based on student economic level.

Facilities

A survey (L. Harris, 2004) of a cross-section of California classroom teachers found that higher-SES, White, and Asian students had greater access to a variety of forms of educational opportunity than their lower-SES, Black, and Asian peers. In particular, the 20% of schools with the most at-risk[46] students were compared to the 50% of schools with the least at-risk students. The 20% of schools with the highest concentrations of African American, Latino, and American Indian[47] students were also compared with the 20% of schools with the lowest concentrations of these groups. About half of the teachers in schools with the highest proportions of at-risk students and teachers in the schools with the most underrepresented minorities reported the physical condition of their schools to be only fair or poor, compared to 34% of teachers in the schools with the fewest at-risk students, and 28% in the schools

[46] In this survey, at-risk was defined as eligibile for free or reduced-price meals, English language learner, or in a family enrolled in CalWorks.

[47] African American, Latino, and American Indian students are referred to as "underrepresented minorities" in this and some other studies (L. Harris, 2004).

with the least underrepresented minorities. Thirty-six percent of teachers in schools serving large populations of underrepresented minorities reported evidence of cockroaches, rats, or mice in the school, compared to 20% in the schools serving few underrepresented minorities. This 16% gap is matched by and correlated with a 13% gap between schools serving the most and the least at-risk students. Little national evidence on the quality of facilities is available, and the effect this has on student learning is likewise uninvestigated, but there may be some effect and it may tend to widen economic and ethnic test score gaps.

Instructional Materials

The California survey (L. Harris, 2004) showed a similar divide with regard to textbooks. Teachers in the schools with more low-income, Black, and Hispanic students reported a lower level of access to textbooks for their students and lower quality for the textbooks that they had. Nationally, NAEP results from 2000 showed that while 78% of teachers of White eighth-grade mathematics students reported having all or most of the resources they needed, the corresponding percentages for teachers of Black and Hispanic students were 66% and 73% (Strutchens et al., 2004). Furthermore, an analysis of NAEP results found that teachers' identification of resource sufficiency was strongly linked to test scores (Grissmer et al., 2000). The study found, in fact, that increasing teacher resource sufficiency would be, for most states, the most cost-effective way to raise test scores.[48]

[48] This finding is congruent with a finding, based on a review of international education production function literature, that inputs not directly valued by teachers (such as books and instructional materials) have effects (in economic terms, marginal products) 10 to 100 times higher than that of input valued by teachers (such as salaries) (Pritchett & Filmer, 1999).

Curricular Offerings

Schools with less low-income students may offer a more advanced curriculum, leading to an increased test score gap between schools. For example, information from the California Basic Data System shows that in that state, high schools with larger populations of disadvantaged students have, along with many other educational disadvantages, less access to Advanced Placement Courses (Betts et al., 2000).

In *Inequality of Access to Educational Resources*, a two-level analysis of the 1992 NAEP trial state assessment of grade 8 mathematics proficiency by Raudenbush, Fotiu, and Cheong (1998), four key educational resources were shown to be more available to White, Asian, and full-price lunch students than to their peers. This was shown to be part of the reason for test score gaps. One of the four resources was access to Algebra in grade 8. The researchers found that African American, Hispanic American, and Native American students were less likely than European American or Asian American students to attend middle schools that offered algebra, as were students whose parents had a lower level of education. In addition, they found an interesting two-level result. Within schools, they unsurprisingly found that students taking algebra scored more highly than those who do not. These students are selected into algebra because of their ability and also presumably benefit from the more advanced instruction. Between schools, in a model containing both individual student course-taking and school offerings, they found a significant test score disadvantage for students attending a school that offered algebra. Such a model effectively compares non-algebra-taking students at schools that offer algebra with non-algebra-taking students at schools that do not, and finds that those at the school not offering algebra do better. They infer that some of the more advanced students in the schools not offering algebra would have been benefiting from an algebra course if it had been offered.

In addition to clarifying the advantages of attending a school that offers algebra, the researchers demonstrated the ways that multi-level modeling can lead to more revealing results than the more common single-level models. More recent NAEP results (Strutchens et al., 2004), though lacking the two-level modeling, show that in the year 2000, Black and Hispanic students were still less likely to take Algebra in grade 8 than their White peers.

Emphasis on Reasoning

The ethnic and economic composition of a school is strongly predictive of the general educational approach taken in a school. Qualitative studies show that, across subject areas, schools with more White and middle- to upper-class students tend to expect more reasoning, while schools with more Black and working-class students tend to emphasize rote learning (Anyon, 1995; Kozol, 2005; 1998). This may have an effect on school average test scores. While the findings of these researchers may be universal, it is also possible that they vary across disciplines. The "scientifically-based" reading pedagogies that low-performing schools are pressed to adopt are uniformly based on lower-level skills such as phonics (Allington, 2003), but a pro-active stance by mathematics educators put them in the forefront of the standards movement with a document that emphasized reasoning for all students (National Council of Teachers of Mathematics, 1989). The rest of this section will focus on equality of access to reasoning and other elements of reform-based pedagogy in mathematics classrooms. The impact of difference in such access is discussed.

The *Inequality of Access* study (Raudenbush et al., 1998), described in the previous section, found that eighth-grade European American and Asian American students were much more likely to be assigned to mathematics teachers who emphasized reasoning than were their Hispanic, Black, and American Indian peers. The education level of their parents also

predicted access to such teachers. The same study showed that being in a classroom that emphasized reasoning was a strong predictor of mathematics proficiency, even controlling for student background. Raudenbush and his colleagues concluded that this difference in teaching styles was a mediator for ethnic test score gaps.

A pair of articles recently published in the *Education Policy Analysis Archives* used more complete measures for teacher classroom practices. Both used two-level models and grade 4 data from 13,511 students on the 2000 NAEP mathematics assessment to address (among other things) the possibility that differential access to a large collection of high-quality pedagogical practices such as those advocated by the National Council of Teachers of Mathematics (1989; 2000) were a part of the explanation for ethnic test-score gaps. These studies are worth reviewing in some detail because of the light they shed on the relationships between pedagogy, ethnicity, socioeconomic level, and test scores and because of their methodological similarities to the study described in this dissertation.

The first of the articles (Wenglinsky, 2004) noted that the emergence of large-scale surveys with large collections of teaching practice variables (e.g., the National Educational Longitudinal Study and NAEP) allowed an advance over prior production function research, which had to rely on the roughest measures of teaching quality. He cited two prior studies (Cohen & Hill, 2000; Wenglinsky, 2002) that suggested a positive impact of teaching for higher-order thinking skills (i.e., reasoning) on mathematics test scores, and clearly drew the important distinction between within-school and between-school test score gaps. He proposed to ask whether instructional practices affect the racial achievement gap more between schools or within schools. He also proposed to find the instructional practices that were most effective for reducing the achievement gap.

Wenglinsky's first between-school model found that, controlling for SES, the expected mean fourth grade NAEP mathematics test score for a school with no Black or Hispanic students was 193. If that school were 100% Black, the expectation would be 27 points (2 to 3 grade levels) lower, or 166. If the school were 100% Hispanic, the expected mean test score would be 16 points (1 to 2 grade levels) lower – 177. The model presumed that schools with more mixed populations would have scores somewhere between the high of 193 and the low of 166.

Similar results in other studies have been explained in four ways. The first explanation is that the skills and dispositions that Black and Hispanic students bring with them to school prepare them less well for schooling as it exists than the skills and dispositions of their White and Asian peers. The prevalence of Spanish speaking in the homes of Hispanic students would be one example of differential student characteristics. The second explanation is that teachers and administrators in any given school are typically less effective at educating Black and Hispanic children than White and Asian children. The third explanation is that the tests themselves are culturally biased. All three of these factors may play a role in creating the test score gaps. Taken together, they could be quantified by looking at ethnic gaps in test scores within schools.

Wenglinsky did this as the within-school part of his multi-level model, finding that on average, within any given school, White and Asian students tended to score 16 points higher than Black students on the grade 4 NAEP mathematics assessment and 8 points higher than Latinos. That these are significantly smaller than the between-school gaps of 27 points and 16 points suggests that part (11 points for Blacks, 8 points for Hispanics) of the explanation of ethnic test-score differences in mathematics at grade 4 lies between schools and is most

susceptible to a fourth explanation: schools that serve large numbers of Black or Hispanic fourth grade students are simply less effective overall than schools that serve lower proportions of these populations.[49]

Wenglinsky then repeated this model with a significant change: he added 3 teacher characteristics (teacher experience, teacher major, and teacher degree) and 20 questions about teacher practices to the equations that predicted school mean grade 4 NAEP mathematics test scores and also to the equations that predicted the within-school slopes (i.e., the ethnic test score gaps within schools). In a brief two paragraphs of analysis of the results, he drew breath-taking conclusions from this seriously flawed model.

> The coefficients for African American and Latino schools are not substantially different from those in the first HLM.[50] This indicates that the introduction of instructional variables does not mitigate the advantage of predominantly white schools over predominantly African American or predominantly Latino ones. The coefficient for African American students within schools is substantially lower than the analogous coefficient from the first HLM (9 points rather than 16 points). Indeed, the coefficient drops to the level of statistical insignificance. The Latino coefficient also changes substantially (from -9 to 27) and is statistically non-significant. Thus, by including the 20 instructional practices, the second HLM can explain away the entire within-school racial gap. (Wenglinsky, 2004, pp. 16-17)

Of the 69 new parameters estimated[51] in the final model, only 9 were found to be significant at the 0.10 level. It is likely that some or all of these significant results were due to chance, yet Wenglinsky chose to overanalyze them with statements like "testing had a

---

[49] This analysis of his results is an elaboration of Wenglinsky's. His conclusion is simpler, but less precise. "The largest gap is between majority black and majority white schools, and the smallest between Latino and white students within the same school" (Wenglinsky, 2004).

[50] HLM, designed by Raudenbush, among others, is the tool used by Wenglinsky for this two-level model.

[51] Wenglinsky adds 20 teacher practices and 3 teacher characteristics, all apparently averaged at the school level, to the prediction of school test score mean, and another 46 predictors to the two equations explaining differences between schools in within-school test score gaps.

disproportionately negative impact on black students, six points above and beyond the three points for all students" (Wenglinsky, 2004, p. 16).

Wenglinsky's final model is deeply flawed (Reardon, 2005), partially because of the addition of 20 instructional variables into an unstructured regression. This "kitchen sink" approach is greatly hampered by the extensive collinearity between the included variables. In the within-school model, race became an non-significant predictor when instructional practices and teacher characteristics are added. Wenglinsky concluded that "by emphasizing certain forms of instruction, school administrators can indeed succeed at closing the racial achievement gap in their schools" (p. 17). But the declining significance level of race is a result of ballooning standard errors, not decreasing effect sizes. A larger sample size or better methodological choices might have improved Wenglinsky's ability to appropriately make the kinds of claims he chose to make. Some methodological improvements might have been (a) removing non-significant practices from the final model; (b) collecting the practice variables into a larger overarching construct, thereby using their collinearity to more reliably define an overarching construct such as "reform teaching" (Mayer, 1999); or (c) investigating practices one at a time. The last approach was used in a study that will be examined in some detail in the next few paragraphs (S. T. Lubienski, 2006).

Partially as a response to Wenglinsky's article, Lubienski (2006) reported on a similar investigation. This study is described in some detail here as an exemplar of the equation: within-school effect + composition effect = between-school effect. Lubienski found no effect for school ethnic composition, but some effect for school SES composition. These findings were, however, just a backdrop for her main study. Like Wenglinsky, she was interested in relationships between instructional practices and achievement, both overall and for particular

subgroups. She characterized her results as "far less definitive and optimistic" (p. 3) than Wenglinsky's because of a difference in framing and "a difference in the care with which findings were interpreted" (p. 3).

She noted rising NAEP mathematics scores and a debate about whether the increase is due to or in spite of increased use of NCTM *Standards*-based reforms in instructional methods. She proposed to offer "a 'bird's-eye' view of the distribution of some reform-oriented instructional methods, and their correlations with achievement for various student groups" (S. T. Lubienski, 2006, p. 3). Such an approach is ideal to address the question of whether the (mal)distribution of reasoning-oriented instructional methods across ethnically and socioeconomically defined groups of students can provide a part of the explanation for test score gaps.

After a succinct description of the "reform-oriented instruction" advocated by the National Council of Teachers of Mathematics, she noted continued debate about the effectiveness of such instruction in improving test scores, and yet more debate about the possibility that the reform-oriented instruction is more helpful to some groups than to others. She summarized her approach: "By exploring the relationship between particular instructional practices and achievement using hierarchical linear models that include both race and SES, this study examines the extent to which race-related achievement gaps that persist after controlling for SES may be related to differences in students' access to particular mathematics instructional practices, as measured by NAEP" (p. 5).

She reported previous NAEP findings on the distribution of reform-oriented mathematics instruction, summarizing that "overall, previous analyses of NAEP data have indicated some potentially important ways in which White, higher-SES students are experiencing more of the

fundamental instructional shifts called for by NCTM than less privileged students" (S. T. Lubienski, 2006, p. 3). She cited the Raudenbush et al. and Wenglinsky studies mentioned previously in this section but noted that the single instructional practice variable of the former study (emphasis on reasoning) was no longer significantly related to race in the 2000 NAEP data and reiterated the methodological weakness of some of Wenglinsky's conclusions in the latter study.

Lubienski broke down her investigation into three questions:

> First, the study examines the extent to which reform-oriented instructional practices are reaching all students, regardless of race. Second, the study investigates whether particular reform-oriented instructional practices correlate positively or negatively with mathematics achievement, after controlling for race, SES, and other potentially confounding variables. Finally, the study considers whether reform-oriented practices correlate similarly with achievement for diverse student groups, regardless of student race or SES. Taken together, these questions probe whether inequities in access to reform-oriented instruction might contribute to achievement gaps, with a particular focus on Black-White and Hispanic-White gaps that persist after controlling for student- and school-level SES. (p. 7)

Lubienski uses a research-grounded factor analysis to create six measures of reform-oriented practice from the 31 potential measures she found in the NAEP teacher surveys. Six clusters resulted: calculator use, facts and skills, collaborative problem-solving, non-number curricular emphasis, writing about mathematics, and manipulative use. Confirmatory factor analyses of these six factors were at least moderately successful. Reasonably strong agreement with a similar set of factor analyses using grade 8 NAEP data also validated the meaningfulness of the latent variables. Because they didn't fit particularly well into any of the other constructs, teacher emphasis on reasoning, use of multiple choice assessments, and teachers' knowledge of the NCTM *Standards* were each kept as observed variables. Each of the latent and observed measures of reform practices was recoded as necessary to make it

either a standardized[52] distribution or binary. In each case, a higher number indicated a greater alignment with teaching practices recommended by the NCTM *Standards* (National Council of Teachers of Mathematics, 1989, 1991, 2000).

In her descriptive analysis, Lubienski found that in the year 2000, teachers of Black and Hispanic students did not report less use of reform-oriented practices than the teachers of their White peers. They did, however, report more use of multiple choice tests. Lubienski's reform-orientation composites were added, one at a time, to the within-school model including ethnicity, SES, gender, and disability at the within-school level, with controls for SES, ethnicity, and school sector at the between-school level. Of the six composite variables, two were found to be significantly[53] related to mathematics achievement in this well-controlled model[54]. For the observed variables, one of the five tests of significance of the observed variables was significant. Lubienski concluded that "collaborative problem solving, teacher knowledge of the NCTM *Standards*, and having a non-number curricular emphasis were all significant, positive predictors of fourth-grade achievement" (p. 18). She found further that the introduction of these variables – alone, all together, or in a variety of combinations – did not diminish the relationships between ethnicity and test scores, concluding that "disparities in reform-oriented instruction, as measured in these models by the teacher-reported NAEP data, do not help explain much of the race-related achievement

[52] Standardized distributions are normal with a mean of 0 and a standard error of 1.

[53] Lubienski required a significance level of .05, reporting levels of .01 and .001 separately.

[54] The structural equation modeling framework would have allowed Lubienski to incorporate an estimate of the measurement error associated with each composite into the model. Instead, the composites were incorrectly assumed to be measured without error, as are all predictors in traditional regression models.

gaps"[55] (p. 19). A look at interactions between the effects of various practices, SES, and ethnicity yielded few significant effects.

To summarize this subsection, there is some evidence that involuntary minority and lower-SES students are less likely to receive forms of instruction that are most strongly advocated by mathematics education leaders, but these pedagogical differences may be declining with time. The evidence is mixed about the role that modifying teaching styles in this way can have on test scores and test score gaps. No recent quantitative research is available on the relationship of school ethnic or economic composition with mathematics teaching practices. Further research in these areas would be valuable. Finally, a well-designed study (Mayer, 1999) of the reliability and validity of the kind of teacher reform-mathematics self-report survey data provided by NAEP and used in the studies described in this subsection (S. T. Lubienski, 2006; Raudenbush et al., 1998; Wenglinsky, 2004) suggests that such data is reliable[56] when used as a composite (as Lubienski did for most of her constructs), but not when used as single observed indicators (as Lubienski did with some variables, and Raudenbush et al. and Wenglinsky did with all variables). Composites of survey self-report about reform practice are also somewhat valid when compared with observational data. A small qualitative comparison (Mayer, 1999) showed the limits of this: two teachers who have been trained in reform methods can report reform practices, but sometimes the implementations vary widely, from excellent to poor. We must ask whether, regardless of

---

[55] Lubienski used her reform-oriented practice variables only in the within-schools model. Wenglinsky's, on the other hand, were all placed in an aggregate form in the between-schools model. Lubienski's approach would tend to underestimate overall effects, Wenglinsky's to overestimate. The ideal approach would be to place the variables in both models to allow a comparison of their individual-level and aggregate-level effects. No such study has been done to date.

[56] Results are similar when teachers rate themselves once and then again a few weeks later.

teaching styles, students in schools with less low-income and involuntary minority students are more likely to have higher-quality teachers and if this difference has a significant impact on test scores.

Teacher Quality

Teacher quality has been measured in many ways, but the evidence is quite clear that schools with students from higher economic levels and with students from Asian and White ethnic backgrounds tend to have higher quality teachers and that those higher quality teachers help produce higher test scores for all of the students in those schools. As described below, many studies have provided evidence for pieces of this argument. Few studies have brought all the pieces together to show how teacher quality partially mediates relationships between school ethnic and economic composition and test scores.

Twenty-one studies have been identified that speak to this mediating role of teacher quality. They are not equivocal. One two-level regression with multiple strong controls shows that within schools, eighth grade students whose mathematics teachers have mathematics majors tend to score more highly on NAEP and that Black and Hispanic students are less likely to have access to such qualified teachers (Raudenbush et al., 1998).

Another five studies document disparities in teacher quality.

1) Schools whose students have higher average SES have teachers with more education, experience, and credentials (Betts et al., 2000; J. Lee & Wong, 2004).
2) Schools with more low-income students lose more of their best teachers (Krei, 2000).

3) Schools with fewer Black and Hispanic students have higher quality teachers as defined using a variety of measures (K. M. Borman et al., 2004; Esch et al., 2005; Teranishi et al., 2004).
4) White teachers are more likely to leave schools with large numbers of Black students than those with less Black students (Freeman, Scafidi, & Sjoquist, 2005).
5) California schools with more low income, Black, Hispanic, and Indian students have fewer qualified teachers (Oakes, Rogers et al., 2004).
6) New York schools with more low income and non-white students have weaker teachers on a wide variety of quality measures (Lankford, Loeb, & Wyckoff, 2002).
7) Low-income and high-minority schools have lower-paid teachers (Education Trust - West, 2005).

Five studies demonstrate a link between teacher quality measures and student test scores.

1) Regression analyses showed that teacher quality was a strong predictor of state test passing percentages in Massachusetts and South Carolina (Darling-Hammond, 2004).
2) Teachers who lack preparation in either subject matter or teaching methods are significantly less effective at producing learning gains as measured by longitudinal test results (Goldhaber & Brewer, 1997).
3) After controlling for student poverty, teacher experience and preparation significantly predict student test scores (Goe, 2002).

4) An analysis of 1996 NAEP data found that the students of teachers with full certification and a major in the field in which they are teaching have higher average NAEP mathematics test scores (Raudenbush et al., 1998).
5) Fully-certified teachers in Houston outperformed those with temporary certificates (Darling-Hammond, 2005).

Three studies (Betts et al., 2000; R. F. Ferguson, 1991; Goe, 2002) help piece together the full mediation path. Each study found that at an aggregate level, schools with less low-SES, semi-voluntary, or involuntary minority students tended to have higher quality teachers and that this fact partially explained test score gaps between groups. In a 2000 California study (Betts et al., 2000), lower-SES schools had more beginning teachers, more teachers with less than a bachelor's degree, and more teachers who were not fully certified. A regression analysis, controlling for student SES, found these resource factors to have a small effect on student test scores. A 2002 California study (Goe, 2002) using a database from the state Department of Education found that schools with large percentages of poor and minority students had more emergency permit teachers and more beginning teachers. A regression analysis found that emergency permit and beginning teachers were less effective at producing high test scores than their better prepared peers.

A well-controlled regression analysis of Texas data (R. F. Ferguson, 1991) found that "differences in the quality of school account for between one quarter and one third of the variation among Texas school districts in students' scores on statewide standardized reading exams. Most of the estimated effect of schooling is due to a single measure of teacher quality: teachers' performance on a statewide recertification exam required of all Texas teachers in 1986" (p. 466). Furthermore, "a primary cause of inequity across districts in the

quality of education is that districts of higher average socioeconomic status find it easier, with any given salary scale, to attract teachers with strong skills and experience" (p. 466). Taken together, the studies mentioned here suggest strongly that schools with more low income or involuntary minority students may have lower-quality faculties, which lead to lower test scores for all students in the schools.

Cultural Dissonance

Many researchers believe that understanding cultural difference is a key to the understanding of the functioning of U.S. classrooms (Delpit, 1995; Gay, 2002; Kochman, 1981; Ladson-Billings, 2001; Rogoff, 2003; Tate, 1995; Zevenbergen, 2000). Well-meaning teachers may use forms of instruction that are not culturally appropriate for their students and may frequently misread student communications (Delpit, 1995, p. 167). Both processes may create unpleasant environments for students and teachers, leading to decreased learning and to disengagement by both parties.

Rogoff's summary of literature on culture and human development (2003) clarifies the centrality of culture to all development. Notions of culture-free learning are a myth, a myth that is particularly strong in the realm of mathematics and particularly damaging to students from poor and non-European backgrounds (Ascher & Ascher, 1997; Bishop, 2000; D'Ambrosio, 1997; Fasheh, 1997; Gerdes, 1997b; Ginsburg, 1986; Lipka, Mohatt, & Ciulistet Group, 1998; Popkewitz, 2004; Powell & Frankenstein, 1997a, 1997c, 1997e; Prediger, 2004; Walkerdine, 1997). For example, because the classic history of mathematics emphasizes European origins and downplays important African, Indian, and Arabic roots (Joseph, 1997; Powell & Frankenstein, 1997d, 1997f), it can be hard for non-European students to see themselves as mathematically able. Additionally, pedagogies designed by and

for the White middle-class may not be culturally appropriate for lower-income students or students from minority groups (Malloy & Malloy, 1998; Walkerdine, 1997). Culturally relevant (Gay, 2002; Gutstein, Lipman, Hernandez, & Reyes, 1997; Ladson-Billings, 1994, 1997; Tate, 1995) and ethnomathematical pedagogies (Anderson, 1997; Bazin & Tamez, 2002; Fasheh, 1997; Gerdes, 1997a; M. Harris, 1997; Lipka et al., 1998; Pinxten, 1997; Powell & Frankenstein, 1997b, 1997e; Zaslavsky, 1997, 1999) attempt to overcome these problems, but such approaches are far from the mainstream of current mathematics instruction.

Teachers are overwhelmingly White, female, and, by virtue of their education and occupational status, middle-class (Whittington, 2002, p. 2). In an autobiographical moment, Walkerdine (1992) describes the reaction of one such teacher to her experiences teaching in an inner city school.

> In 1986 I became a primary school teacher. I was swayed by the romantic promise of progressivism in education, and I linked poverty and inner-city decay with the terrible regimentation and the "old-fashioned" repressive and silencing methods...And four o'clock found me frequently sobbing quietly at my desk, behind the shut door where none of the old, strict teachers, who didn't like my ways, could see me. (p. 15)

What Walkerdine believes she experienced was a common disconnect between romantic notions of humanistic and individualistic teaching and the economic, cultural, and social realities of the educational system. She blames a progressive education that makes powerlessness, the product of oppression, invisible. She calls teachers "guardians of an impossible dream, reason's dream of democratic harmony" (Walkerdine, 1992, p. 22).

Delpit's ethnographic interviews with twelve educators of color (1995, pp. 105-127) illuminate the problem of cultural dissonance from another perspective. These teachers do not suffer from the kind of naïveté described by Walkerdine. They are, as a rule, well aware

of at least some of the structural inequalities built into our nation's economic and educational systems. But they face a different, perhaps even more demoralizing challenge – that of being the interface between dominated students and a dominating system. Most of the teachers Delpit interviewed believe that accounts of their own experiences were not validated in teacher education programs or in their subsequent teaching lives. One Native Alaskan teacher who completed teacher education but never entered the teaching profession said, "I began to think I must be a radical or a racist or something because *they* always said, 'Everything's great, why make a fuss' I'd say, 'No. It's not!'" (p. 109).

One Black man left teaching after two years in an Alabama junior high school

> ...because I got totally dissatisfied with the system I was a part of. The staff was 98 percent white and 2 percent black. Near the end of the first year, I realized that I was the only staff member interested in helping *students* progress, not in just covering the course material. (pp. 110-111)

A female Native Alaskan teacher said that in her experience, teachers always separated themselves from villagers. "I'd have to choose sides – either with the teachers or with the village – and I'd choose the village. It would be too hard being in the middle like that" (p. 111). Delpit's interviews suggest that cultural conflict in schools is more important than most White, middle-class educators understand. Gary Howard's little book *We Can't Teach What We Don't Know* (1999) provides a roadmap for the difficult personal journey involved in coming to this kind of deep recognition from the experience of being a White male teacher. In the meantime, teachers in schools with large involuntary minority or low-income student bodies may continue to face cultural dissonance – which can result in lowered student test scores and disengagement from the educational enterprise.

Other researchers have provided more fine-grained analyses of the ways that cultural dissonance affects students and teachers. Kochman (1981), for example, provides detailed

analyses of the miscommunications that are everyday occurrences between Blacks and Whites in the United States. Heath's (1982) ethnographic work shows that working-class parent-child interactions tend to be characterized by direct instructions, while middle-class children are more likely to receive indirect instruction. This can lead to cultural conflict in the classroom when teachers give instructions to students that sound like weak requests (see also Delpit, 1995, p. 168). Zevenbergen (2000) examines an Australian Grade 6 testing scheme to show how the language differences between students from dominant and dominated classes create differential abilities to make sense of the specialized vocabulary, semantic structure, and lexical density of mathematics texts. Lubienski's (2000) study of her own classroom suggests that "some characteristics of discussion-intensive mathematics classrooms might be more aligned with middle-class cultures" (p. 377). Stiff (1990, p. 156) provides many examples of "how Black expression in mathematics classrooms produces negative feedback from teachers." Berry's (2002) interviews with mathematically achieving young black men and their families show that many of them were overlooked for academically gifted programs by the schools, or worse, seen as behavior problems. Only the intervention of strong adults in elementary school allowed them to reach their potential in the school. Even in all-Black schools, "messages of black cultural deviance" are transmitted to students (Tyson, 2003, p. 326).

Cultural conflicts of the larger society may cause different groups to have different degrees of faith that the schooling system is working for them. For example, American Indians are aware that the explicit original purpose of schooling for them was to remove them from their culture and thereby to take their culture from them (Spring, 2004). Children were often forcibly taken from their homes, forbidden to speak their native languages, and

encouraged to disassociate themselves from their families. Spring provides similar histories of "deculturalization" for other minority groups in the United States. Knowledge of such injustices may lead students in schools serving the children of involuntary minority groups to be more ambivalent in their attachments to school.

Ogbu's cross-cultural research (1997) compares Black and American Indian students to members of other involuntary, castelike minority groups around the world. Such groups may resist an education system that asks for conformity to culturally foreign norms and offers little true hope for advancement.[57] Such resistance would be a cultural, not just an individual, phenomenon. In one Washington, D.C. area school, for example, Black students discouraged each other from academic effort, calling such efforts "Acting White[58]" (Fordham & Ogbu, 1986). Perry (2003) calls these the "dilemmas of achievement facing African-American students as members of a group subject to an ideology of intellectual and cultural inferiority" (p. vii). Steele and Aronson (1998) show that these dilemmas have direct effect on test scores in the form of stereotype threat. Willis's ethnographic study of a group of White working-class students in England (1981) shows that cultural dissonance can occur on economic as well as ethnic grounds.

To summarize the argument of this subsection, the composition of a student body may have a profound effect on the culture of the school. This could affect students as they engage

---

[57] Some qualitative work shows that segregation increases this sense of economic hopelessness by depriving students of the opportunity of seeing examples of "flesh and blood" people making it (O'Connor, 1998).

[58] The "Acting White" hypothesis is both well-known and controversial. Cook and Ludwig (1998) use data from the National Educational Longitudinal Survey (NELS) to suggest that "Black high school students are not particularly alienated from school." (p. 390) Tyson's (2002) ethnographic study of two all-Black elementary schools suggests that Black elementary students are very much engaged in school and that elements of disengagement begin to form as they experience a lack of success and inappropriately harsh criticisms of their cultural styles (Tyson, 2003).

in their most important developmental task as secondary school students: identity formation. Wexler (1988) described how the characteristics of developing identity or “becoming somebody” vary across high school contexts. For example, students in Black underclass schools tend to focus on "chillin out." In working-class schools, "cranking up spirit" is a form of extra-curricular identity work. The children of the status-conscious professional middle-class work on "mellowing out," while corporate executive children "have fun."

In particular, the further the student body is from the dominant culture of the society, the more the school will suffer from cultural conflict (Bourdieu & Passeron, 1990). The composition of the student body, then, may affect the ability of the school to improve the test scores of all of the students in the school, at least partially because of a tendency of all parties in the school to disengage from the schooling process. One quantitative study provides some evidence in support of this hypothesis. A study of the National Education Longitudinal Survey (NELS) found that schools with higher percentages of minority students had higher absenteeism and lower levels of classroom preparation and participation among Hispanic, Black, and White students alike (Finn & Voelkl, 1993). This suggests that all students in a school are affected by the composition of the school, even those that are not members of involuntary minority groups. This may reflect cultural dissonance and may lead to Full-School Disengagement.

Section 4. Full-School Engagement

The engagements of parents, teachers, students, and administrators in the schooling process are important resources for effective schools. Taken together, these resources are referred to here as Full-School Engagement. A few researchers have investigated the possibility that the ethnic and economic composition of schools affect, through a variety of

mechanisms, various aspects of this Full-School Engagement. These studies will be described in the following paragraphs.

Until 2007, *Education Week*'s influential annual review of state education policies, *Quality Counts* (e.g. McCabe, 2006b), included "school climate" as one of its important quality measures. In 2006, states received grades ranging from *B* to *D*+, based on (a) NAEP principal survey reports of parent involvement in schools and of the degree to which student absenteeism, tardiness, and misbehavior are problems; (b) the existence of state surveys of teachers, parents, and students about school conditions, school facilities, and parent involvement; (c) state efforts to document and improve school safety; (d) state support for charter schools; (e) school size; (f) class size; and (g) state support for facility improvement. The use of school climate as one of only four measures[59] of the state's contribution to equitable and effective education suggests the importance of the concept. The breadth of the categories included as measures of school climate suggests a need to narrow the construct somewhat for theoretical clarity. For this reason, the construct of Full-School Engagement is developed for this study.

A review of the literature on *The Organization of Effective Secondary Schools* (V. E. Lee et al., 1993) suggests that schools with more low-income or involuntary minority students tend to be more bureaucratic; that these bureaucracies discourage engagement by parents, teachers, and students; and that this lack of engagement tends to depress test scores. These results are not seen by the researchers as inevitable. More communitarian approaches

[59] Along with school climate, the other indicators of the quality of state educational programs are: "standards and accountability," "efforts to improve teacher quality," and "resource equity" (McCabe, 2006b, p. 78).

(Comer, 2001; Comer, Michael, Haynes, & Joyner, 1999) can potentially increase Full-School Engagement.

Two NAEP studies used advanced modeling techniques to show a mediating role for "school disciplinary climate" (Raudenbush et al., 1998) and "school social environment" (Wenglinsky, 1997). A previously discussed multi-level model (Raudenbush et al., 1998) showed disciplinary climate to be one of four resources[60] that are advantageous for all categories of eighth grade mathematics students, and more available to White, Asian, and higher SES students. Disciplinary climate was operationalized in this study as a composite based on NAEP principal survey reports of problems in the school: student tardiness and absenteeism, cutting classes, physical conflicts, drug and alcohol use, teacher absenteeism, and race or cultural conflict.

Wenglinsky's structural equation model (1997) built on the production-function research tradition to explore the ways that money makes a difference for education. The "school social environment" scale was a composite created from responses to the 1992 grade 8 NAEP survey of principals. The indicators were: student tardiness, student absenteeism, teacher control over instruction, teacher control over course content, regard for school property, teacher absenteeism, and student class-cutting. An SEM analysis suggested that the socioeconomic composition of a school was directly related to academic achievement and also related to achievement by way of the mediating school environment variable. These two studies suggest that schools with lower SES, American Indian, Black, or Hispanic students may have lower levels of Full-School Engagement and therefore lower test scores. The next

[60] The other three resources are teacher preparation, access to algebra instruction, and teacher reported emphasis on reasoning.

few sections will provide more details about the components of Full-School Engagement. These components are teacher engagement, parent engagement, student engagement, student resistance, and administrative optimism.

Teacher Engagement

There is much evidence to suggest that teachers in schools with more economically and ethnically advantaged students are more engaged in their work. They have higher efficacy, higher morale, less absenteeism, less turnover, and higher expectations for their students. This evidence is presented in the next few paragraphs.

Teacher morale is affected by the engagement of other actors in the educational system and by accessibility of physical resources. A national Teacher of the Year and Pulitzer-Prize winning author described how a move to an elite New York City high school was a turning point in his teaching career, in his autobiographical book *Teacher Man,* in a section called *Coming Alive in Room 205.*

> [My new supervisor] took me to a room where books were organized by grade. It was dazzling to see them ranged on shelves reaching to the twenty-foot ceilings and stacked on carts for delivery to classrooms…If you asked the boys and girls of Stuyvessant High School to write three hundred and fifty words on any subject they might respond with five hundred. They had words to spare. (McCourt, 2005, p. 185)

In New York and many other cities, elite schools like Stuyvessant are starkly different from the vast majority of schools in the system. The ethnic and economic compositions of the student bodies, the resources available to the school, and, as a consequence, the morale of the teachers in the schools differ dramatically (Kozol, 2005). A second exemplar and the results of a strong study confirm the role of teacher morale in giving more academic gifts to the students who already have the most.

Mary Foster, "exactly the type of teacher that experts say low-performing high schools need," left her 11-year career at just such a school to move to one of her district's "premier" high schools. She had been proud to teach at majority-Black Southern High School

> ...because she felt that many of the teachers who criticized the school's academic standing could not have made it there. But she got tired of not feeling supported by the principals who were under immense pressure to boost test scores…little things like having to buy her own notebooks and pens wore on her. Foster could not believe it when she arrived at Riverside and found a closet full of supplies. (Hannah-Jones, 2006)

Foster describes "very emotionally draining work," growing tired of "having to fill the gaps when three and four teachers would quit at one time. I just felt like it was time for a change so that I could continue to teach." She said it improved her morale to be surrounded by other strong teachers and damaged it to see them abandon the school. "Every year things got a little darker and the teachers ended up being that much more demoralized" (Hannah-Jones, 2006).

The examples of Foster and McCourt suggest that teacher morale is often related to the ethnic and economic composition of a school. These suggestions are confirmed by a multi-level analysis (V. E. Lee, Dedrick, & Smith, 1991) of national High School and Beyond data. The researchers find student characteristics, at both the within-school and between-school levels, to be one of the strong predictors of teacher morale. This study of 8,488 students and approximately 10,000 teachers in 354 schools combines four self-reported variables relating to sense of success and overall job satisfaction as a measure of self-efficacy. A preliminary multi-level finding is that teacher efficacy varies both within (89% of total variance) and between schools (11%). Within schools, efficacy is predicted by teacher control over the classroom and by the academic level of the students taught (as compared with the rest of the school). Teacher gender, ethnicity, salary, experience, and major subject taught are all non-

significant. A full between-schools model finds that schools having greater size, less disorder, stronger leadership, a stronger sense of faculty community, and – most relevant to this section – higher aggregate SES, promote greater teacher efficacy. This large-scale study demonstrates that Foster and McCourt are not alone. Together with modifiable school organizational factors, the socioeconomic composition of the student body is an important predictor of teacher efficacy and morale.

Expectations

A robust literature shows that teachers expect less of some groups of students than of others. This section will present evidence that expectations are a key piece of teacher engagement and are strongly related to achievement.

A classic of the expectations literature is Rist's (1970) analysis of a set of students in a 100% black, inner city school. He found that (a) kindergarten teachers have an "ideal type" of the "good" student in mind, a type that is informed by many middle-class prejudices as well as by actual academic ability; (b) teachers group the students on that basis; ((c) the "high" group receives more and higher quality attention; ((d) a "caste" forms within the classroom; and ((e) academic differentials are created and passed on from grade to grade with little upward mobility possible. This study suggests the mechanisms by which differential expectations tend to widen academic gaps between groups of children within schools.

Another classic, Anyon's (1981) qualitative study of elementary schools in New Jersey, suggests that differential expectations by economic level also occur powerfully between schools. She characterized the schools she studied as working-class, middle-class, affluent professional, and executive elite. A male teacher in a working-class school said his was a "tough" school. He'd been nervous to teach there until the principal told him, "Just do your

best. If they learn to add and subtract, that's a bonus. If not, don't worry about it." Another, when asked about important knowledge for her children, said, "Well, we keep them busy" (p. 7). The dominant theme Anyon found in working-class schools was resistance. She noted active and passive resistance by students along with pleasure at angering the teacher (p. 11).

In the middle-class school, knowledge was seen to come from books. Math class included some sense that procedures were meaningful. The teacher usually gave several ways to do a problem and told the students: "I want to make sure you understand what you're doing" (pp. 13-14). Knowledge was seen by the students as worth having, as a way to truly get ahead.

The focus of the professional middle-class school was on helping students learn to think. In mathematics, students were expected to learn through discovery and direct experience with manipulatives and projects. Student requests for help were often responded to with questions like "What do you think?" Knowledge for these students was more about thinking than about something to be found in books. Anyon characterized the dominant theme of this school as extreme individualism.

Anyon concluded that New Jersey's, and by extension, the nation's, schools are designed to effectively reproduce the social class structure of society. The academic expectations teachers have for most of the students mirror the expectations they have for the children's futures, which mirror the present roles of their parents. The most relevant comment for this study came from a second grade teacher in the working-class school. She said she would not want to work in the district's school for the gifted and talented because "you have to work too hard" (p. 7). Low expectations for students translate to low engagement by teachers.

Delpit's interviews with teachers of color are again useful (1995, pp. 105-127). A Native Alaskan woman who had completed teacher education decided not to teach in part because of

her experience as a student teacher – "The teachers … had the attitude that the students were hard to teach. Some told me they didn't think the [Native Alaskan] kids knew how to think" (p. 111). Berry's qualitative studies (Berry III, 2002, 2004a, 2004b) of successful Black male middle-school mathematics students point to many factors leading to achievement, but in most cases the students had to overcome the low expectations of their teachers.

Ferguson (1998) provides the strongest recent summary of the research on teacher expectations and the Black-White test score gap. He finds that teachers have lower expectations for Black students than for White students, and that these expectations affect test scores. The expectations are linked, however, more to student behaviors than directly to student race. Expectations have more effect on Black and low-income children than on middle-class White children. Part of the reason is that while White students tend to work to please their parents, Black students are often much more motivated by pleasing the teacher. As Berry's interviews suggested, high-performing Black students feel that each year they have to prove to their teachers that they are capable of honors-level work.

Ferguson (1998) scans the research and national data in search of explanations for the racially different treatment of students. He finds that, especially in the earliest grades, teachers feel that Black students are less likely than White to care about doing well, to get along with teachers, and to work hard at school. This may lead teachers to withdraw support from those students. He reports on a study by Willis and Brophy (1974). When first-grade teachers were asked to nominate students to attachment, indifference, concern, and rejection groups, only in the rejection group were nonwhite boys overrepresented. Ferguson also reports on a study of Ft. Wayne students in which student explanations of bad grades focused entirely on the racial prejudice of the teachers. He does not believe the expected differences

are inevitable, suggesting instead that "greater responsiveness to individual children can weaken the link between past and future performance" (p. 303). He provides examples of teaching that does not limit students.

The work of Ferguson, Anyon, and many others makes clear the important role of teacher expectations within and between schools in the creation and maintenance of test score gaps. I will end this section with a quote from the passionate Lisa Delpit. "Educators must cease questioning the capacity of low income students of color and, instead, create rigorous, engaging instruction based on knowing who the students are, including their cultural, intellectual, historical, and political legacies" (2003, p. 14). If this transformation were to occur on a national basis, teacher engagement gaps would probably decline, as would test score gaps.

Parent Engagement

"When educators huddle amongst themselves, disappointed, disgruntled, and besieged, they vent their frustration on the deficiencies and intractability of parents" (Redding, 2005, p. 8).

The idea that differences in parent engagement levels help create between-school test score gaps is accepted wisdom among teachers. Overcoming these differences is a central piece of many reform proposals, such as Moses and Cobb's (2001) Algebra Project, which calls on Black parents to demand Algebra as the new civil right, and Comer's School Development Program (Comer et al., 1996), which includes parents heavily in the development of multi-dimensionally nurturing school communities. Parent involvement at the level of "planning, implementing, and evaluating school improvement activities," as in the Comer model, is one of the 11 components that the U.S. Department of Education uses to

define a Comprehensive School Reform Model (G. D. Borman, Hewes, Overman, & Brown, 2003, p. 127). Nevertheless, involving low-income parents of color in school remains a major challenge (A. C. Barton, Drake, Perez, Louis, & George, 2004; Martin, 2000; Porter, 1996; Strutchens, 2000) and federal enthusiasm for the effort is waning (Redding, 2005) as some studies find little evidence linking such efforts to test scores (G. D. Borman et al., 2003; Redding, 2005).

Coleman (1988) and Bourdieu (1990) provide the theoretical foundations for the notion that involvement in schools is one of the ways that privileged parents transmit privilege to their children. Coleman sees some parents as having more "social capital," which they expend at the school to help develop the "human capital" of their child. Parents who are more strongly linked to the school community and the wider community are advantaged because (a) teachers and administrators will feel a sense of obligation and commitment to look after their children; (b) they are more privy to school communication channels; and (c) they are more attuned to the social norms of the school, further fostering effective communication that can give their child an advantage. These advantages translate into improved learning for their children, which develops the human capital and employability of those children.

Bourdieu (1990) focuses more on differences in "cultural capital" than "social capital." He would see U.S. schools as embodying a White middle-class culture that is more open and inviting to the participation of White middle-class parents. Lower-class and non-White parents are disadvantaged by this system in multiple ways: the assessments are essentially tests of White middle-class knowledge; the system privileges White middle-class behaviors from students and from parents. Less advantaged families have cultural capital, but it is often not the kind valued by the school. This can lead to cultural conflict between parents and

school.[61] Such parents may engage with the school less effectively or disengage altogether. Either way, the children of these less advantaged parents are the losers.

Lee and Bowen (2006) studied a variety of forms of parent engagement with an eye to the theories of Coleman and Bourdieu. Their data source was a sample of 415 third to fifth graders in a community bordering a major urban center in the southeastern United States. They included only students whose parents identified as African-American, Hispanic/Latino, or European-American. Forty percent of the children received free or reduced-price lunches at school, 15% were Hispanic/Latino, and 34% were African-American. They found that European-American parents were more likely to be involved at the school, as were parents whose children did not receive free or reduced-price lunch. European-American parents were least likely to be involved with their children's time management, and most likely to have parent-child educational discussions. Non-poor parents were most likely to engage in parent-child discussions and have high educational expectations, but they were least likely to directly manage their children's time. According to teacher ratings, European-American students had the highest achievement, followed by African-American, and then Hispanic/Latino. Non-poor students outperformed recipients of free and reduced-price lunches.

The researchers used multi-level multiple regression to predict academic achievement. A model containing just demographic variables explained 24% of the variance, with school lunch status, parent education, and African-American ethnicity significantly related to achievement. Addition of five parent involvement constructs added nine percentage points to the predictions; this model predicted 33% of achievement variance. Parent involvement at

[61] Cultural conflict can also occur between children and school, as was discussed in a previous section.

school and parent educational expectations were the only two parent involvement constructs to show a significant relation to test scores. Their addition decreased the size of school lunch and African-American coefficients, leading the researchers to conclude that parent involvement at school and parent educational expectations were mediators of the well-known relationships of student ethnicity and economic level with achievement. The researchers continued with interesting interaction term modeling, but the study could have been improved by the use multi-level techniques or mediation modeling.

These investigations of the within-school effects of parental involvement on the achievement of their own children are supplemented by studies of the between-school effects of aggregate parent involvement. A review of the research on effective schools (V. E. Lee et al., 1993) suggests that schools with more overall parent involvement are more effective with all of their students. A review of education production functions (Pritchett & Filmer, 1999) suggests a reason: educational inputs valued by teachers (such as salary) are dramatically overused in comparison to inputs less valued by teachers (such as textbooks). Expenditures not valued by teachers are 10 to 100 times as effective at raising test scores as the teacher-valued expenditures. The researchers suggest that increased parent input into educational decision-making would encourage schools to use their funds more efficiently.

Willms's (2006) international study of PIRLS (the Progress in International Reading Literacy Study) supported the relationship between aggregate parental engagement and school effectiveness. PIRLS assessed the literacy skills of fourth graders in 35 countries around the world, simultaneously surveying students, parents, teachers, and administrators. The multi-level study, paralleling the study of PISA reported in a prior section, found that

parental support[62] for the school's academic mission was one of the few variables consistently linked to average school achievement, adjusting for the characteristics of individual students and other predictor variables at both the within-school and between-school levels.

The research described above supports the importance of parent engagement in encouraging student achievement, but many studies of attempts to actually increase this engagement have had weak results. Martin's (2000) qualitative study of a poor, Black urban community in which Algebra for All had been implemented found that despite the program's emphasis on community participation, parent attitudes remained negative and unhelpful toward mathematics learning, leading to bad test results.

The meta-analysis by Borman et al. (2003) of comprehensive school reform (CSR) models had mixed results. CSR models that included parent engagement components produced lower test scores on average than those that did not, but one of the three programs to demonstrate the most effectiveness, Comer's School Development Program (SDP), had a heavy focus on parent engagement. The other two, Success for All and Direct Instruction, are both based on very structured curricular interventions and are far more expensive than the Comer model, which focuses on the needs of the whole child and involvement of the whole community. It seems that while interventions focusing directly on preparation for tests can be effective at raising test scores, well-implemented whole-child, whole-school interventions like Comer's School Development Program might improve test scores and much more. As Noblit noted in

[62] For the PIRLS study, parental support was based on principals' overall assessment of parental support for student achievement, as well as their assessment of the percentage of students whose parents would (a) volunteer to assist in classrooms or elsewhere in the school, (b) attend parent-teacher conferences, (c) attend sporting, social, and cultural events at the school, or (d) fundraise or otherwise support the school.

concluding his study of Comer schools: "To improve achievement, a school does need to focus on students, curricula, and instruction, but to improve a school it is necessary to change its structure and culture" (2001, p. 132). This requires strong parent involvement, but the literature on effective ways to encourage this involvement is still weak.

A strong contribution to the literature on the encouragement of parent engagement comes from an integrative review of the psychological literature by Hoover-Dempsey and Sandler (1997). They find that the key elements in a parent's decision to participate in school activities are (a) the parent's role construction; (b) the parents' sense of efficacy for helping their children succeed in school; and (c) general invitations, demands, and opportunities for involvement. No parent involvement program will be successful without addressing all three elements. As will be seen, each of the elements is related to parent economic and ethnic background.

The researchers (Hoover-Dempsey & Sandler, 1997) report on ethnic and economic differences in role construction. There are ethnic differences in the answer to the question of who should initiate parent involvement in the school. White parents tend to take that role, while other ethnic groups tend to leave the initiative with the teacher. Working-class parents tend to have a more "separated" view of home and school. They get the children ready and send them to the school. By contrast, middle-class parents are more inclined to have an "interconnected" view of school that involves active monitoring of their children in and out of school, and even intervention in school decisions when needed. These patterns are by no means universal. Some working-class parents favor a more interconnected view and some middle-class parents stay at some distance from the process, but overall, the pattern has an effect on differential parent engagement (Hoover-Dempsey & Sandler, 1997).

Even if parents view engagement in their child's schooling as a part of their role, they may fail to become involved if they do not feel a strong sense of efficacy in helping their children succeed in school. Parents with higher self-efficacy set higher standards for themselves and their children and put forth greater effort to achieve their goals because they believe they are attainable. Parents with more education have higher self-efficacy for helping their child succeed in school and are therefore more likely to participate in school activities. This participation has positive results for student achievement (Hoover-Dempsey & Sandler, 1997).

Parent attribution of success and failure is an element of self-efficacy. These attributions have a cultural component. For example, Chinese parents are more likely than U.S. parents to attribute success to hard work. U.S. parents are more likely to consider luck and ability the key factors for success. Such attributions make parents less willing to spend their time in supporting their child academically (Hoover-Dempsey & Sandler, 1997).

The final, most concrete, most used, and least important factor influencing parental decisions about involvement in the schools is general opportunities, invitations, and demands presented by children, schools, and teachers (Hoover-Dempsey & Sandler, 1997). Like the other two factors, this one tends to favor more advantaged students. For example, in a representative national survey, Spanish-speaking families report receiving fewer communications about their children than non-Spanish-speaking families[63] (National Center for Education Statistics, 2006).

---

[63] The Spanish-speaking families report receiving fewer personal notes or emails about their children; fewer newsletters, memos, or notices addressed to all parents; and fewer invitations to general meetings, parent-teacher conferences, school or class events, and volunteer opportunities.

Hoover-Dempsey and Sandler (1997) summarized that all efforts to improve parent engagement should explicitly address role construction, efficacy, and invitation if they are to be successful – "particularly among parents whose experiences have resulted in relatively weak role construction or efficacy" (p. 36). Their framework also provides some of the explanation for differential parent engagement in segregated schools serving different populations. This differential engagement has effects on student achievement (Willms, 2006). Differential parent engagement may be one of the explanations for gaps in scores between schools with large percentages of low-income involuntary minority students and those with fewer students in these categories.

Student Engagement

Differential student engagement may help to create between-school test score gaps. Specifically, schools with more White, Asian, and well-to-do students may be less likely to have problems with students missing or being late to school and less likely to report problems with the achievement attitudes of their students. These and other student engagement factors may be part of the reason for test score gaps.

Ogbu (1992) proposes that involuntary minority groups tend to develop an oppositional culture that creates a pattern of disengagement from school. The quantitative research following on his theory has been tangled by what Mickelson (1990) calls "the attitude-achievement" paradox. Black students often express *more* positive academic attitudes and aspirations than their peers, but those attitudes are not translated into more positive school behaviors such as attendance and time spent on homework (Ainsworth-Darnell & Downey, 1998, p. 536; Blau, 2003, p. 208; M. K. Johnson, Crosnoe, & Elder Jr., 2001; Ogbu, 1988; Ogbu & Simons, 1994).

A survey of 2,245 Oakland minority students (Ogbu & Simons, 1994) generally supports Ogbu's claims about voluntary and involuntary minorities. The researchers compare African-American, Mexican/Latino, and Chinese-American students, expecting to find involuntary minority African-Americans resistant to schooling, voluntary Chinese-Americans compliant, and semi-voluntary Mexican-American/Latino students in the middle. The Chinese-American students fit the pattern:

> They are willing to conform to the dominant society's norms in order to succeed and do not fear that crossing cultural boundaries will harm their social identity. Their educational strategies involve conforming to the expectations that schools have of good students. They have high aspirations, work hard in and out of school, and conform to teachers' behavioral expectations and as a result they succeed. (p. 20)

By contrast, the educational model for involuntary African-Americans is found to be

> ...ambivalent….On the one hand, they report their parents and community believe in education as the route to making it in society. At the same time they are sensitive to prejudice and discrimination and believe equally in non-educational sources of knowledge. This produces ambivalent educational strategies which involve claims of parental support and high aspirations among both students and parents and exaggerated claims of school success. At the same time, they report less effort than the Chinese Americans. Further, they report that school success is stigmatized by students in general not by their close friends. This suggests they are ambivalent about crossing cultural boundaries which they perceive school success to require, for fear of displacing their social identity. These contradictions in their beliefs and stated behavior may in the context of substandard schools where they are the objects of low expectations make it difficult to provide the effort necessary for school success. (p. 20)

The pattern found for Mexican-American/Latinos is more complex. Their sensitivity to barriers to success is, as expected because of their semi-involuntary status, less than that of African-Americans but greater than that of Chinese-Americans. The same is true of the stigma associated with doing well in school. They share with both groups a stated faith in education as a means to success, but are midway between African-Americans and Chinese-

Americans in their valuing of street knowledge and desire for sport or entertainment success. Mexican-Americans show the lowest educational aspirations of the three groups.

Blau (2003, pp. 97-132) provides further data to support Ogbu's ethnic distinctions. Using student responses to the national urban High School Effectiveness Study (HSES), she creates five Getting in Trouble (GIT) scales and compares the weighted factor scores of Asian, Black, Latino, and White students. Asians have low scores on all five GIT scales; Blacks and Latinos have high scores on the two scales most directly related to engagement in school. Latinos score highest on the Cutting Classes scale, which encompasses self-reported frequency of lateness to school, cutting, and skipping of classes. Black students score second highest. Black students score by far the highest on the Unprepared for Class scale, which encompasses coming to class without pencil and paper, books, and homework done. Latinos score second highest. These ethnic patterns are maintained even controlling for student gender, poverty status, and family structure. Students from poor families score higher than those from non-poor families on the Unprepared for Class scale, but not on the Cutting Classes scale.

Finn and Voelkl (1993) use the National Education Longitudinal Survey (NELS 88) to define five "engagement indicators" (p. 256). ABS-TARDY incorporates teacher reports of whether a youngster was frequently absent from class or tardy. NOT-ENGAGED uses teacher reports of student inattention, missing homework, and inattention/disruption in class. ATTENDANCE is based on students' reports of the number of times they missed school, skipped classes, or arrived late, and of the number of times their parents were contacted about attendance problems. PREPARATION is a self-report of coming to class without pencil and paper, without books, and without completed homework. BEHAVIOR is based on

students' reports of being sent to the principal's office for misbehaving, parental warnings about their child's behavior, and fights with another student.

Using multi-level modeling and controlling for a wide range of other factors in a NELS sub-sample of at-risk students, they find student-level SES to be predictive of each category of student engagement behavior. Student ethnicity is not related to these behaviors, but the percentage of minority students in the school is. At the .05 level, controlling for the ethnicity of individual students, schools with more minorities are more likely to have problems with absences, tardiness, attendance, and disengaged students.

However, it may be hypothesized that school economic and ethnic compositions are related to the school engagement of students, which has an effect on test score outcomes. This thesis seems particularly solid when student engagement is measured by active participation in schooling, rather than by attitudes.

Student Resistance

"High concentrations of disadvantaged students can adversely affect the school's ability to maintain the social order and can foment peer cultures that act in opposition to the school's academic aims" (V. E. Lee et al., 1993, p. 180). This summary of a review of research on effective schools can serve as a summary of this section. In particular, Black and working-class student resistance to the schooling enterprise is documented. Order is shown to be a resource for schools. Disorderly and sometimes dangerous schooling is shown to be particularly characteristic of schools with large minority populations. The role of student resistance as a mediator of ethnic and economic composition effects is shown. All claims are documented with qualitative or quantitative research.

Student resistance can be seen as the flip side of student engagement, but the two constructs are well-viewed separately. Student resistance activities tend to be active and aggressive as opposed to the passive approaches of the disengaged student. They are, of course, frequently seen together. In fact, high resistance and low engagement are often exhibited by the same students.

Paul Willis's qualitative study of "the Lads," a group of White working-class British students whose only interest in school is to have a "laff" before entering the working world of their parents, provides a classic picture of student resistance. Fights, especially racial conflict, and violations of the rules were a key part of their approach to the schooling process. Those boys who followed school rules were denigrated as "ear 'oles" (P. Willis, 1981). Solomon (1988) and MacLeod (1995) describe similar patterns among Black Canadian and White New England working-class youth, respectively. A recent ethnography shows how Black boys are socialized into the role of "bad boys" in the nation's elementary schools.

> For example, one day a fifth-grade African American boy who was always in trouble saw the file folder with his name on the desk. "I got a lot in there, don't I? Who else got one that big?" he asked. There was awe in his voice at his accomplishment. He had made an important mark on the school. (A. A. Ferguson, 2000, p. 9)

The quantitative studies reported in the previous section (student engagement) include information about student resistance. The construct of student resistance is similar to Finn and Voelkl's (1993) BEHAVIOR variable, which was shown to be related to student SES, but not to individual or composite school ethnicity. Ogbu and Simons (1994) found suspension rates to be far higher for African-American students than Hispanic or Asian students (p. 10), relating these behaviors to the lack of trust Black students feel for the school institution. This lack of trust is mirrored in the finding of Johnson et al. (2001) that African-American

students report the lowest attachment to school of the three ethnic groups they studied. Blau found that Blacks, Whites, and poor students are more likely to report having had discipline problems than Latinos, Asians, and non-poor students (p. 104). In summary, Black and poor students appear to be particularly likely to engage in resistance to the process of schooling for some of the same reasons that they are among the most likely students to disengage from the schooling process.

According to the effective schools literature, order and discipline are among the most important characteristics of schools that promote strong learning (Finn & Voelkl, 1993, p. 254). Willms's (2006) look at PIRLS and PISA shows this to be true on an international level. School disciplinary climate and student-teacher relations are significant predictors of school mean test scores, even with a strong collection of within-school and between-school covariates. It stands to reason, then, that the differential student resistance associated with student ethnic and class identity would be reflected in school mean test scores.

Black and Hispanic students are more likely than White students to fear for their safety in school and out of school. Nine percent of Black students and 10 percent of Hispanic students, but only 4 percent of White students, reported that they were afraid of being attacked at school (Dinkes, Cataldi, Kena, & Baum, 2006). According to data from the 2003-2004 School Survey on Crime and Safety, students in schools with more than 50% minority students are more likely than students in schools with more White students to

- be involved in a violent incident (43.4 incidents per thousand students over the course of the school year at high-minority schools)
- be involved in a seriously violent incident (1.8)
- be threatened with physical attack with a weapon (0.6)

- be threatened with physical attack without a weapon (16.2)
- be robbed with a weapon (0.1)
- be robbed without a weapon (0.4)
- commit vandalism (4.8)
- be involved in a hate crime (0.2)
- be involved in a gang-related crime (1.2)

(Guerino, Hurwitz, Noonan, & Kaffenberger, 2006).

Raudenbush, Fotiu, and Cheong's (1998) multi-level NAEP study shows "school disciplinary climate" to be one of the resources that influences school effectiveness and is less available to Native Americans, Hispanic-Americans, African-Americans, and lower-SES students. The evidence is strong that student resistance mediates ethnic and economic composition effects on school mean test scores.

Administrative Optimism

An optimistic principal probably improves school effectiveness. This kind of principal is more likely to see solutions than problems and to be engaged in implementing those solutions. Harris and Willomer (1998) provide support for this thesis, with an important caveat about measurement. In a survey of teachers, they find that teacher perceptions of their principal's optimism are correlated to their sense of school effectiveness. Principal self-report of optimism is not so correlated, suggesting that this variable is best measured indirectly.

No other high-quality research has been found on the topic of Administrative Optimism, partly because of a lack of high-quality research on any aspect of school administration. Although research reviews found hundreds of articles, few were actual research reports, only a small number were published in scholarly journals, and the vast majority suffer from major

methodological flaws (V. E. Lee et al., 1993). The current study, then, will be among the first to measure this potentially important construct.

Full-School Engagement and School Climate

The positive engagement of any of the parties to the schooling process is a powerful motivator for the engagement of each of the other parties. Louis and Smith report that parent engagement generates teacher engagement (1992); Brewster and Bowen's (2004) survey finds that engaged teachers of Latino students create engaged students. Jenkins's (1995) path model shows higher parental engagement improving student commitment, which in turn decreases student criminal activity, misbehavior, and nonattendance. Tucker and colleagues surveyed 117 Black elementary and secondary students (2002) and found that teacher involvement exerts a strong and direct effect on student engagement even when controlling for grade level and self-system variables. An autobiographical Teacher of the Year (McCourt, 2005) described how engaged students helped bring him "alive." Ferguson (1998) found that student behaviors affect teacher expectations. Silins and Mulford (2004) performed a path analysis on 96 schools and found support for a complex web of relationships among 12 constructs, including resources available to help staff, leadership, valuing of staff, community focus, students' perception of teachers' work, and student participation. The examples of relationships among these various forms of engagement are multitudinous, but I am aware of no research that has operationalized them together as a single emergent full-school construct.

The closest construct in the literature is school climate, but operationalization of school climate has varied widely. Some versions overlap partially with Full-School Engagement (e.g. Borkan, Capa, Figueiredo, & Loadman, 2003; M. K. Johnson et al., 2001; Raudenbush et al., 1998; Raudenbush, Rowan, & Kang, 1991), others hardly at all (e.g. Buckley, Storino,

& Sebastiani, 2003; Gregoire & Algina, 2000; Shindler, Taylor, Cadenas, & Jones, 2003). A School Climate Inventory (SCI) is used to assess "impacts of reform initiatives in relation to 7 dimensions logically and empirically linked with factors associated with effective school organizational climates" (Ross & Lowther, 2003, pp. 222-223). The dimensions of the inventory are closely related to Full-School Engagement. They are listed below with the related construct from Full-School Engagement in parentheses.

1. *Order.* Orderliness of environment and appropriateness of student behaviors (Student Resistance).

2. *Leadership.* Extent to which administration provides instructional leadership (Administrative Optimism).

3. *Environment.* The extent to which positive learning environments exist (Teacher Engagement, Student Engagement).

4. *Involvement.* The extent to which parents and the community are involved in the school (Parent Engagement).

5. *Instruction.* The extent to which the instructional program is well developed and implemented (Teacher Engagement).

6. *Expectations.* The extent to which students are expected to learn and be responsible (Teacher Engagement).

7. *Collaboration.* The extent to which the administration, faculty, and students cooperate and participate in problem solving (Student Engagement, Teacher Engagement).

The complex relationships among the various parties to schooling are reciprocal and self-reinforcing to such a degree that they may be most parsimoniously viewed as a single second-order construct called Full-School Engagement. It is similar to the construct

measured by The School Climate Inventory, but a new term is used because the construct is only tangentially related to many other operationalizations of school climate.

Section 5. Full-School Engagement as a Mediator of Composition Effects

The next few sections investigate the role of Full-School Engagement as a mediator of ethnic and economic composition effects on test scores. It is shown that school ethnic and economic composition may predict Full-School Engagement, that Full-School Engagement may predict adjusted school mean mathematics test scores, and therefore, that Full-School Engagement may partially mediate the effect of school ethnic and economic composition on adjusted school mean test scores.

School Composition May Predict Full-School Engagement

Evidence has been provided in previous sections that school economic or ethnic composition may be predictive of Student Engagement, Teacher Engagement, Parent Engagement, and Student Resistance. Taken together, these constructs represent Full-School Engagement. It is therefore logical to assume that school economic or ethnic composition may be predictive of Full-School Engagement. Some quantitative literature on the organization of effective schools provides more insight into this question (Bryk & Driscoll, 1988; V. E. Lee et al., 1993; V. E. Lee & Smith, 1995).

The key distinction made in this literature on the organization of effective schools is between “communal” and “bureaucratic” organization of schools. An early article (Bryk & Driscoll, 1988) developed an index of communal school organization using the national High School and Beyond survey. A communal school is “a social organization consisting of cooperative adults who share a common purpose and where daily life for both adults and students is informed by shared values and a common agenda of activities. The positive

relationship between parents and school staff provides important support for school aims" (Abstract). A communal school has a system of shared values, a common agenda of activities, and a distinctive pattern of social relations. HSB data showed that communal organization, higher in Catholic schools and small schools, improves teacher efficacy, teacher enjoyment, and staff morale while decreasing absenteeism. It decreases student class-cutting, absenteeism, and classroom disorder, while increasing interest in academics and mathematics achievement. The researchers did not, however, find the expected link between communitarian organization and school ethnic or economic makeup.

Five years later, a review of the research on effective schools found "sufficient evidence for us to conclude that aspects of student composition influence organizational operations and these features, in turn, affect both teachers and students" (V. E. Lee et al., 1993, p. 180). They found ethnographic evidence that

> ...in an important sense, the communitarian ethos typical of the smaller and more homogeneous public high schools of the 1950s, albeit often discriminatory and intolerant as a result of closure to outsiders, was shattered by legal desegregation efforts. The resultant increase in student diversity and the problems arising from it in individual schools was accommodated by a variety of bureaucratic mechanisms. The restructuring of curriculum and related efforts to repair social relations with the school, however, resulted in a systemic departure from previous strong institutional norms promoting academic achievement for all students. (V. E. Lee et al., 1993, p. 180)

An analysis of 11,794 sophomores in 830 high schools from the first two waves of the National Educational Longitudinal Study (Lee & Smith, 1995) found that schools with lower average-student SES are more likely to be bureaucratically organized than those with higher average SES, and that schools with a higher percentage of minority students are also more likely to be bureaucratically organized – and consequently to have lower levels of Full-School Engagement. More research is needed, but it seems likely that school ethnic and economic composition are predictive of Full-School Engagement.

Full-School Engagement May Predict Adjusted School-Mean Test Scores

The NELS study reported above (V. E. Lee & Smith, 1995) found strong multi-level results favoring communal organization. Such organization was associated with higher test scores in a variety of subjects, as well as higher student engagement. These results replicate those found in the earlier HSB study (Bryk & Driscoll, 1988).

This evidence, which focuses on the way that school organization affects test scores, provides a supplement to the wealth of evidence provided in earlier sections of the effects that teacher engagement, parent engagement, student engagement, and student resistance have on test scores. It seems likely that Full-School Engagement is related to school mean mathematics test scores, even with adjustments made for within-school predictors such as student ethnicity and economic level.

Full-School Engagement May Partially Mediate Composition Effects

The study that comes closest to examining the role of Full-School Engagement as a partial mediator of ethnic and economic composition effects on test scores is Wenglinsky's 1997 analysis of data from NAEP, the Common Core of Data, and the Teacher's Cost Index, all available from the National Center for Educational Statistics. As previously described, this structural equation model places "School Environment" in a mediating position between school socioeconomic composition and 1992 grade 8 school average NAEP mathematics test scores. The model is accepted, allowing acceptance of the hypothesis implicitly made by the model: "School Environment" partially mediates the effect of school socioeconomic composition on grade 8 school average NAEP mathematics test score. This School Environment variable is a smaller version of the Full-School Engagement variable to be operationalized in this study. It includes

- student tardiness
- student absenteeism
- teacher control over instruction
- teacher control over course content
- regard for school property
- teacher absenteeism
- student class cutting

Section 6. Use of the National Assessment of Educational Progress for this Study

NAEP provides a good tool for the testing of this theory because of its psychometric quality, its background survey, and its educational significance. NAEP is the only representative national survey of what students know and can do in mathematics and other school subjects. The 2003 survey provides a very large sample size and uses oversampling methods that allow for inferences to smaller groups, such as American Indians / Alaskan Natives, that were too poorly represented in previous surveys for safe inference.

The Main NAEP mathematics exam, designed in collaboration with the National Council of Teachers of Mathematics (Lindquist, 2001), is better able to measure the advanced competencies required of today's mathematics students than many other standardized tests. In addition, NAEP's design represents the highest psychometric standards (see Horkay, 1999 for details). Another advantage of NAEP is that students and school administrators are both surveyed, allowing connections to be made between scores and background characteristics. Finally, NAEP is known as The Nation's Report Card (National Center for Education Statistics, 2005) for a reason. NCLB legislation requires that state testing programs be

validated with NAEP results (Miller, 2003), thereby moving NAEP in the direction of a de facto national test. Although there are no stakes for the young people taking the test, there is no test in the nation with higher stakes for our educational system. It is clearly worthy of study and well suited to the testing of the theory proposed in this study.

Section 7. Two-Level Structural Equation Modeling of NAEP Mathematics Scores

Access to the full NAEP database requires strict adherence to a tight security regime in order to protect the privacy of students and educators. The NAEP survey's complex sampling structure requires careful handling or special software. Only in recent years has software (HLM) been available to facilitate the calculation of multi-level models (Raudenbush, Bryk, Cheong, & Congdon, 2000). Even more recently, single-step modeling of two-level structural equation models has been facilitated (L. K. Muthén & Muthén, 1998-2005). Advances in computing power have also increased researchers' abilities to estimate complex models. Even with these software and hardware advances, the challenges involved in multi-level and structural equation modeling with NAEP data are great. For this reason, very few researchers have done so, despite the benefits of the approach. I have found only one study that estimates a multi-level structural equation model with NAEP data. A strong understanding of its strengths and weaknesses provides a guide to the methodology section that follows.

Wenglinsky (2002) used grade 8 NAEP data from the 1996 assessment to estimate a two-level structural equation model using structural equation modeling software with a preprocessor designed to create the two levels of the model. His focus was on the between-

school model; within schools he simply allowed student SES[64] to covary with student mathematics achievement score.[65] Between schools, his final, parsimonious[66] model included three measures of basic school resources,[67] three measures of professional development,[68] and six measures of teaching practice.[69] Wenglinsky's structural equation model proposed that school mean test scores, adjusted for student-level SES, were affected by all of these factors. Because he used structural equation modeling, he was able to use composites, corrected for error[70], to measure these predictor variables more reliably than his regression-bound predecessors could do. In addition, he was able to model the relationships between the predictor variables. In his model, teaching practices mediated the relationships between professional development and school mean test scores. Both teaching practices and

[64] Wenglinsky's SES variable was a factor measured by mother's educational level; father's educational level; and family access to newspapers, encylopedias, magazines, and books.

[65] Because no student takes a complete assessment, NAEP researchers are not provided with a single accurate test score for any student. They are provided instead with five "plausible values" for the student's test score. Wenglinsky followed standard NAEP procedure by estimating each model five times, using a different NAEP plausible value each time, averaging the parameter results and increasing the standard errors.

[66] Wenglinsky followed the common statistical practice of including a wide variety of variables in his model, then removing those that prove insignificant in the context of all the competing variables. The advantage of this approach is parsimony – conceptual simplicity. Its disadvantage is that it may lead to inappropriate removal of some variables from the model because their effect is masked by the effect of a related variable or variables (Hedges & Greenwald, 1996).

[67] School resources included in Wenglinsky's parsimonious model were represented by an SES variable that is an aggregate of the SES values for the assessed students in that school, class size as reported by the teacher, and teacher major.

[68] The three measures of professional development included in the parsimonious model were factors representing professional development for diversity, professional development for higher order thinking, and the amount of time spent on professional development. All observed variables were means of the responses of the sampled teachers within the school.

[69] The six measures of teacher practice were means of the responses of the sampled teachers to their use of lower-order instruction, higher-order instruction, hands-on instruction, authentic assessment, and traditional assessment.

[70] Wenglinsky explains that multilevel structural equation models "take measurement error into account in two ways. For one, the factor models explicitly measure the amount of variance in the latent variables unexplained by the manifest variables. In additino factor models can actually reduce measurement error by generating latent variables from multiple manifest variables." (Wenglinsky, 2002, p. 9)

professional development mediated the effects of school resources on school mean test scores. He found that variables at all three levels had an effect of school mean test scores and claimed that "the effects of classroom practices, when added to those of other teacher characteristics, are comparable in size to those of student background, suggesting that teachers can contribute as much to student learning as the students themselves" (p. 1).

This study provides the only available model of the use of two-level structural equation modeling of NAEP mathematics data, but it is flawed in some respects. By ignoring race and ethnicity, Wenglinsky misses an important set of student background and school composition factors. At the within-school level, SES is clearly a cause of student achievement, not merely correlated. At the between-school level, school mean SES should, perhaps, also be modeled as a cause of class size and teacher major. It seems unlikely that the cause would be reversed. Finally, and most importantly, Wenglinsky confuses the meaning of his results. He claims to compare teacher factors with student SES, but he has controlled at the within-school level for student SES. He is therefore comparing teacher factors with school SES composition. A more appropriate summary would be: "the effects of classroom practices, when added to those of other teacher characteristics, are comparable in size to those of the SES composition of the school." This much more limited claim would not allow the inference that "teachers can contribute as much to student learning as students themselves" (p. 1).

Section 8. Summary

With evidence ranging from solid to light, this literature review provides theoretical and empirical support for the models that will be proposed in Chapter 3. This is a critically important part of structural equation modeling because many models can always be proposed to represent any social science situation. It is much stronger to use a dataset to provide

confirmatory or disconfirmatory evidence of a theory derived from prior theory and study than to explore without such guidance. The guidance has been provided. The next chapter will present and explain the resulting models.

CHAPTER THREE - METHODOLOGY

Chapter 3 describes the methodology of the current study. Section 1 provides a framework for the methods, including (a) the purpose of the study and the questions it addresses, (b) an introduction to structural equation modeling methods, (c) a brief description of the five models used to answer the questions, (d) an overview of the grade 8 NAEP 2003 mathematics database, and (e) a description of structural equation modeling as it is practiced in this study. Section 2 provides a detailed overview of the study's five models. Section 3 addresses technical issues and software choice.

Section 1. Framework of the Study

Purpose and Questions

This study concerns Full-School Engagement, that is, the degree to which an entire school community actively engages in the academic mission of the school. The specific purpose of this study is to investigate the degree to which Full-School Engagement explains grade 8 mathematics test score gaps that exist between economically or ethnically differing groups of students. The following four questions are addressed:

Question 1. Can a single second-order latent variable called Full-School Engagement measure a constellation of factors representing administrative, parent, teacher, and student engagement in the academic mission of a school?

Question 2. Do the economic or ethnic compositions of a school predict that school's mean grade 8 mathematics tests scores, adjusted for the ethnicity and

economic level of the individual students in that school (i.e., composition effects)?

Question 3. What are the relationships that exist among the economic and ethnic compositions of a school, Full-School Engagement, and adjusted school mean grade 8 mathematics test scores?

Question 4. Does Full-School Engagement mediate any of the composition effects identified in Question 2?

Structural Equation Modeling

These questions are investigated using the scientific method. Clear hypotheses based on prior scientific study are generated. These hypotheses are then tested using observational data. In the case of this study, the hypotheses are specified as complex causal mathematical models. These models are analyzed using structural equation modeling (SEM), a method that provides a particularly powerful set of tools to specify, test, and estimate mathematical models of the relationships that exist between sets of real-world variables. SEM is able to model with both observed and latent variables, and to include estimates of measurement error within the modeling framework.

SEM is sometimes called covariance structure modeling because the observational data being modeled take the form of a *covariance matrix*. *Covariance*[71] is a measure of the degree to which a pair of observed variables varies together. A *covariance matrix* contains the

[71] Covariance is strongly related to the more commonly understood concept of correlation. A correlation of 0 indicates that the value of one variable has no implication for the value of the other. If two variables have a correlation of 1 or -1, on the other hand, then the value of one of the variables in a given record would absolutely determine the value of the other variable in that record. Variables may be highly correlated because one causes the other or because another variable or set of variables causes both. The only difference between covariance and correlation is that covariance values can go above 1 and below -1 because the values are multiplied by the individual variances of each variable.

covariances of each pair of observed variables in a model. The covariance of a variable with itself is the *variance* of that variable. These variances are found on the main diagonal of a covariance matrix.[72]

SEM presumes that the covariances found in the covariance matrix are the result of causal processes, other forms of unexplained covariance, and error. The researcher specifies hypothesized causal relationships between the variables in a *path model.* Additional, unobserved *latent variables* are added to this path model if they are justified by prior knowledge and measurable by other variables that can be included in the model. Most latent variables are measured by observed variables. These are called first-order latent variables. Some latent variables include first-order latent variables as "measures." These are called second-order latent variables.[73] A model of the relationships among observed and latent variables is specified, based on prior research and theory. The model generally includes *error* and *disturbance* terms because no model can precisely predict the variables in a system.

Structural equation models can be specified in three forms – diagrams, equations, or series of computer program statements. The models can also be described in English sentences, but the other three formats are more concise and precise. Diagrams are the most user-friendly of the formats. Observed variables are represented as rectangles; latent variables, which are only indirectly measured, are represented as ovals. Arrows indicate causal relationships. Equations

---

[72] Because a covariance matrix contains variances on the diagonal, it is often referred to as a variance-covariance matrix, but this may be regarded as redundant because the covariance of a variable with itself is always equal to its variance.

[73] Figure 2 provides a good example of first- and second-order latent variables in the context of a measurement model. Student Engagement, Student Resistance, Teacher Engagement, Parent Engagement, and Administrative Optimism are all first-order latent variables because they are measured by the 23 observed variables represented by rectangles. Full-School Engagement is a second-order latent variable because it is measured by the set of five first-order latent variables.

(generally in matrix form) are the most precise format. Each SEM-estimating computer program has its own statement syntax. No matter how it is presented, each model includes some number of *parameters* to be estimated by the software.

The researcher provides the software with a model specification and a dataset. In some cases, the dataset is simply the covariance matrix. In other cases, the researcher provides raw data and the software calculates the observed covariance matrix. Through an iterative process, the software estimates values for the parameters that minimize the differences between the *observed* covariance matrix and the *model-implied* covariance matrix resulting from the model specification and parameter estimates[74].

The basic hypothesis of any structural equation model is that a set of values for the parameters can be found such that the population covariance matrix of observed variables is equal to the model-implied covariance matrix. A model is said to be *underidentified* when there is insufficient information in the covariance matrix to uniquely estimate model parameters. Similarly, a model may be *just-identified*[75] or *over-identified* when exactly enough or more than enough needed information, respectively, is available to estimate parameters.

A variety of indices is used to measure the degree to which the two matrices vary from perfect fit in the sample. These are called *measures of overall fit*. A model for which the measures of overall fit are good and the parameter estimates are reasonable is considered a good model. In cases of good fit, the researcher may proceed to investigate specific

---

[74] In some cases, such as when missing data is treated by the model, variable means are included along with covariances as data for the model.

[75] Traditional regression models are always just identified. They have exactly as many parameters as information in the covariance matrix. They can be almost always be solved by closed-form, non-iterative methods. Their fit is always perfect.

parameter estimates – first, to see if they are reasonable as a secondary measure of overall model fit and second, to draw tentative conclusions about the magnitudes of the relationships in the model. One can never, however, be certain that the model is correct. As with any model, the best we can do is to see if the data are consistent with the model.

Five Models

Five key models are specified and estimated to answer this study's questions. Each is presented first as a verbal hypothesis and later as a more detailed graphic model. A full description of the model and its notation is provided at that point.

Model 1. **Full-School Engagement CFA Model** (Question 1). The first hypothesis tested by this model is that Full-School Engagement is well described as a second-order latent variable incorporating five first-order latent variables (Student Resistance, Student Engagement, Teacher Engagement, Parent Engagement, and Administrative Optimism). This hypothesis is tested using a confirmatory factor analysis (CFA) structural equation model. The model is a single-level, between-school model. No information about individual students is provided because it uses the schools dataset.

Model 2. **Baseline Regression Model** (Question 2). It is hypothesized that student ethnicity and economic level are related to grade 8 mathematics test scores. A simple regression analysis of grade 8 NAEP mathematics test scores on individual economic and ethnic variables tests this hypothesis. This model replicates the kind of simple single-level analyses that have been done many times in the past. As with these analyses, no attempt to distinguish the within-school and between-school parts of the relationships is made, despite the

clustering of students within schools. This model is one of three associated with the second question of the study. The second and third models are designed to differentiate the within-school and between-school parts of this relationship.

Model 3. **Baseline two-level model** (Question 2). Variance in grade 8 mathematics test scores is hypothesized to occur both between and within schools. A two-level, no predictor model allows the separation of test score variance into between-school and within-school components. This is the second model associated with the second question in the study.

Model 4. **Composition effects model** (Question 2). The hypothesis tested by this two-level model with predictors at two levels is that school economic and ethnic compositions each affect school adjusted mean test scores. The adjustments (also known as controls) are for the ethnicity and economic level of individual students. This is the final model used to answer the second question in the study.

Model 5. **Mediation model** (Questions 3 and 4). The final model adds the Full-School Engagement latent variable to Model 4, testing the hypothesis that Full-School Engagement partially mediates the effects of school economic or ethnic compositions on grade 8 adjusted school mean mathematics test scores. This model addresses the third question of the study and, when compared with Model 4, is used to addresses the fourth and final question.

These five models are diagrammed and described in this chapter. Implementation details are included as needed. Recall, however, that SEM requires both models and data. The data used for this study are described next.

The Grade 8 NAEP 2003 Mathematics Database

This study's models are tested using the restricted-access Grade 8 NAEP 2003 Mathematics database. NAEP is the only nationally representative test of what students know and can do in mathematics and other subjects. A large stratified random sample of students (Appendix C contains details of the sampling strategy) from across the country were assessed in mathematics in the spring of 2003 (National Center for Education Statistics, 2003, p. 20). Each sampled student completed a survey, as did that student's mathematics teacher and an administrator from each school. These brief surveys allow for analyses of some of the factors that may affect achievement on this important examination. The 2003 grade 8 mathematics NAEP restricted-use database is provided to users as two datasets – a *schools dataset* with one record for each of 6,334 targeted schools, and a *students dataset* with one record for each of 162,727 targeted students (Rogers & Stoeckel, 2004, p. 56). Records from the schools dataset can be added to the appropriate records of the student dataset to create a *combined dataset* containing all associated information in the *schools dataset* and the *students dataset.*

Since the year 2000, all states have been required to participate in NAEP, increasing the sample size dramatically and allowing for meaningful analyses of multiple subpopulations at both the state and national levels. All 50 states and 3 jurisdictions[76] participated and met

[76] Along with the 50 states, the District of Columbia, the Department of Defense Domestic Dependent Elementary and Secondary Schools, and the Department of Defense Dependents Schools (overseas) participated in the 2003 Mathematics NAEP Assessment (National Center for Education Statistics, 2003, p. 23).

minimum guidelines for reporting their results in 2003. The multi-stage, clustered sample design includes oversampling of certain groups of schools (such as private schools and schools with large populations of American Indians) to provide sufficient power for analyses of these smaller groups. Weights are provided with the sample. When these weights are included in analyses, the results are representative of all grade 8 students in the nation. All regions of the nation, all ethnicities, all economic levels, and public and private school students are included. Appendix C contains more details about the NAEP sampling strategy, including an illustrative figure.

The large size of the NAEP datasets permits this study to confirm and reconfirm the proposed models. For initial confirmatory analyses and possible respecifications, a randomly chosen quarter of each dataset was used. Final analyses of key models were conducted with either (a) the remaining three-quarters of the appropriate dataset or (b) the entire dataset. Model 1, designed to measure Full-School Engagement, uses only school-level variables and is therefore estimated with the schools dataset alone. The first estimation and modifications are made using a random subsample of one quarter of the schools dataset. A slightly respecified model is replicated on the complete schools dataset. Models 2 through 5 use both student-level and school-level variables. Because the students dataset does not include information about schools, the two datasets are combined for these analyses into a dataset with 162,727 records and all relevant student- and school-level variables. The school-level variables in this combined dataset are identical for every student in a given school. Models 2 through 5 are estimated first with a new random subsample containing one quarter of the records. Replication results are then obtained with the remaining three-quarters of the combined dataset.

Structural Equation Modeling as Practiced in this Study

Each of the five models proposed in this study is estimated with SEM. One approach to SEM is described here.

(1) Based on a review of the literature, **specify** a model of the relationships that may exist between a set of variables – using equations, matrices, a path diagram, and program code. Hypothesize that this model provides a good approximation of reality.

(2) Confirm that the model is **identified** (unique values of parameters can be found), using rules defined in the statistical literature (e.g. Bollen, 1989).

(3) Provide data and model (usually in the form of program code) to a software program designed to **estimate** the model using an iterative computational method such as maximum likelihood. The program estimates optimal values for the model's parameters (regression-style coefficients, factor loadings, measurement error estimates, other variance estimates, and, in the case of the advanced model studied here, threshold estimates for the transformation of categorical variables to their underlying normal distributions). Optimal values are those that come closest to defining a model-implied covariance matrix that matches the observed covariance matrix.

(4) Compare the model-implied covariance matrix with the observed covariance matrix to gain a sense of the **overall fit** of the proposed model with the data.

(5) **Respecify** the model as needed to improve its overall fit with the data and prior scientific evidence.

(6) If an adequate fit is found, **interpret** the model and its parameter estimates; if not, reject the hypothesis that the model provides a good approximation of reality.

For each of the five proposed models, the first two steps of this process are completed as much as possible, and the rest of the steps are defined. Boldface type is used in each section to emphasize this six-step process. The estimations, fit characterizations, respecification as needed, and interpretation are all discussed in Chapter 4. Equations, matrices, and more detail are provided in Appendix A.

Section 2. Overview of the Five Models

Section 2 provides a detailed overview of the five models used to answer the questions posed for this study. Each model is described and specified in graphical form. As new variables and constructs are introduced, they are described in the text. As new technical ideas are introduced, they are discussed. The theoretical identification of each model is considered and the steps that will be detailed in Chapter 4 (estimation, consideration of fit, respecification as needed, and interpretation) are introduced for each model. The equations and matrices corresponding to each model are presented in Appendix A.

Model 1: Confirmatory Factor Analysis of Full-School Engagement

Model 1 is a confirmatory factor analysis designed to answer Question 1: *Can a single second-order latent variable called Full-School Engagement measure a constellation of factors representing administrative, parent, teacher, and student engagement in the academic mission of a school?* The purpose of confirmatory factor analysis is to confirm that a set of observed variables is well predicted by a set of latent variables (also known as *factors* or *constructs*) in a manner specified before estimation by the researcher. The model **specified** in Figure 2 is single-level because it concerns schools only, not students. It is second-order because the Full-School Engagement latent variable is measured by other latent variables.

Therefore, it is a single-level, second-order CFA. CFA models can be estimated using structural equation modeling (SEM) or other techniques. One advantage of SEM for CFAs is that the measurement model can later be combined with other latent and observed variables in a path model (Bollen, 1989, pp. 313-315; L. K. Muthén & Muthén, 1998-2005, pp. 52-53). Another is that the assumptions of the model are graphically depicted, highlighting assumptions sometimes left implicit by other statistical techniques.

As shown in Figure 2, the hypothesis of this model is that Full-School Engagement can be appropriately viewed as a second-order factor, with effects on five first-order factors: Student Engagement, Student Resistance, Teacher Engagement, Parent Engagement, and Administrative Optimism. The five first-order factors are measured by 23 variables, all of which are taken from the NAEP administrative questionnaire completed at each school site by the principal or a designee. These factors and their related measures are discussed in Chapter 2, described in some detail in Appendix B, and defined here.

- Student Engagement is the degree to which students are motivated to perform academic activities. It is measured by administrative reports of the percentage of students absent on a given day, of student attitudes toward achievement, and of problems with student absenteeism and tardiness.
- Student Resistance is the degree to which an adversarial relationship exists between teachers and a significant subgroup of students. It includes seven measures of active student behaviors that serve, intentionally or not, to disrupt the school process for all students. The first six are administrative reports of problems with physical conflict among students, racial or cultural conflicts, gang activities, student misbehavior in class, physical conflict between students and teachers, and

vandalism. The final measure for Student Resistance is the administrator's assessment of student regard for school property.

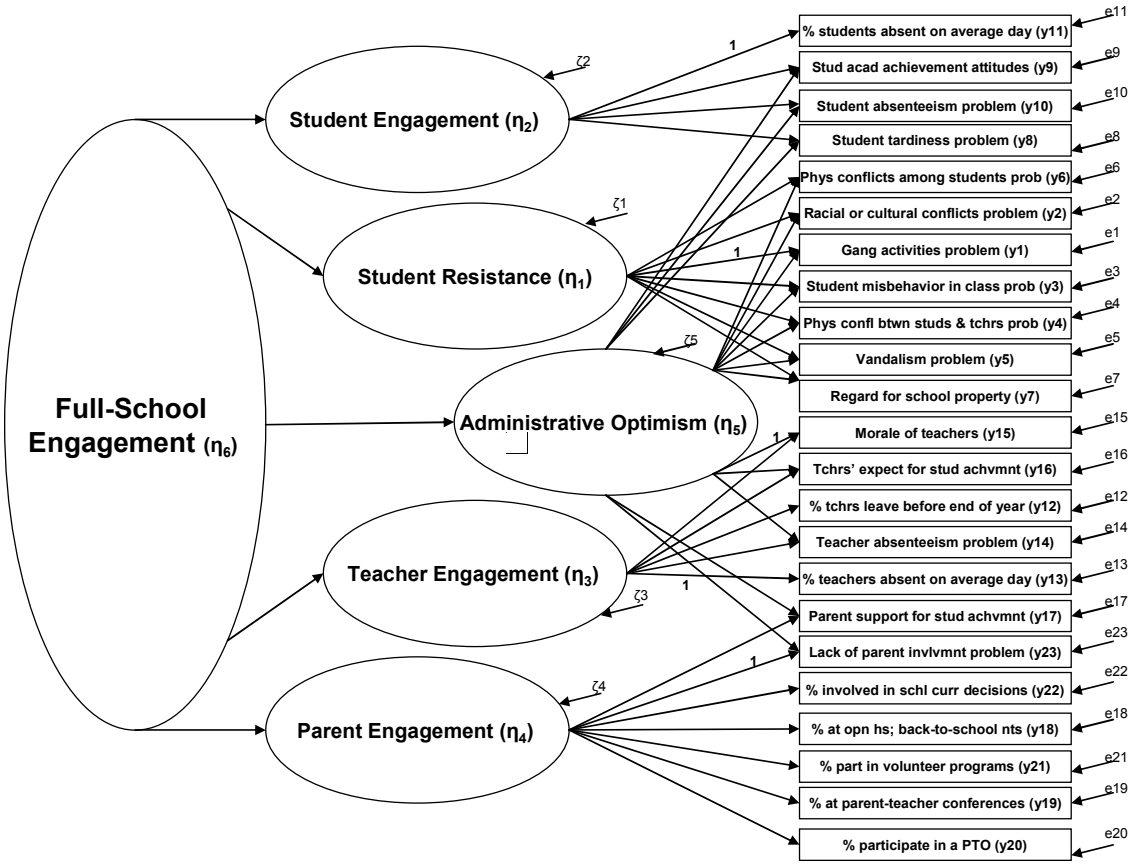
- Teacher Engagement, the extent that teachers are involved in the schooling process, is measured by administrative assessment of morale of teachers, teacher expectations for student success, the percentage of teachers that typically leave before the end of a school year, the degree to which teacher absenteeism is a problem, and the percentage of teachers absent on an average day.
- Parent Engagement is also seen as key to the schooling process. Parents are expected to support their child's achievement, to be involved generally, to be involved in school curriculum decisions, to attend open houses and back-to-school nights, to volunteer, to attend parent-teacher conferences, and to participate in the PTO. Each of these variables, as assessed through a survey completed by an administrator, is included as part of the Parent Engagement latent variable.
- Administrative Optimism, the extent that administrators view the other parties as engaged, is addressed below.

All 23 variables measuring the four forms of engagement come from a NAEP survey typically completed by a single administrator. The more subjective of the variables are taken to measure administrative optimism along with the domain about which the question is being asked. For example, an optimistic administrator at one school is more likely to report positive morale of teachers than a pessimistic administrator at another school, even if an objective outside observer saw the same level of morale at the two schools. This model suggests that administrative optimism is the degree to which the administrator who completes the survey

views students, parents, and teachers as engaged or resistant, controlling for the degree to which they actually are engaged or resistant (as measured by less subjective variables).

Finally, the model presumes that an overarching second-order latent variable called Full-School Engagement has a significant effect on each of the five first-order factors. The hypothesis is that there is something important and measurable about a school that affects the engagement of each party in the schooling process.

Figure 2. Model 1. Confirmatory Factor Analysis of Full-School Engagement – Initial Specification



Equations, matrices, and assumptions representing this model are presented in Appendix A. As previously discussed, structural equation models are attempts to explain the variances and covariances found in a set of observed variables. In this case, we have 23 observed variables; each variable takes different values from school to school. The model presented in

Figure 2 presents a series of strong assumptions about exactly how the 23 observed variables vary and covary, most importantly in the form of proposed causal paths and missing causal paths. If the assumptions are correct or nearly correct, a computer program will be able to assign values to each of the causal paths and other parameters such that the covariance matrix computed based on the assumptions and parameters will resemble closely the sample observed variable covariance matrix.

The sample observed variable covariance matrix is a matrix of the variances and covariances found among the 23 variables in the administrator's survey; neither structure nor causal relationship is imposed on it. The model diagrammed in Figure 2, on the other hand, is a literature-based attempt to describe the structural relationships that exist among those 23 variables. Using SEM techniques, I will be able to test whether this proposed model fits the sample covariance matrix.

The model in Figure 2 proposes that the variance observed in each of 23 variables has two or three causes. These causal relationships are indicated by arrows. For example, administrative reports of the morale of teachers are presumed to be caused by Teacher Engagement (as indicated by the arrow from Teacher Engagement to the observed variable), Administrative Optimism (as indicated by the arrow from Administrative Optimism to the observed variable), and a *residual* (indicated by the arrow from e15 to the observed variable).

Teacher Engagement is a latent variable. It is not directly observed; instead it is measured by way of its effects. By hypothesizing that Teacher Engagement affects exactly 5 of the 23 observed variables, the claim is made that these 5 variables will covary more strongly with each other than with the other 18 observed variables. If this is true, then the viability of the Teacher Engagement construct is supported.

According to the model, Teacher Engagement is one of the causes of the reported level of *morale of teachers*. The emphasis on *reported* level suggests that the optimism of the administrator completing the survey may also affect that variable, as reflected by the arrow from Administrative Optimism to the *morale of teachers* observed variable in Figure 2. That all 23 observed variables are taken from the same survey and many of them are prone to this kind of subjective evaluation can be viewed a methodological problem. However, it is taken instead as an opportunity to permit measurement of the Administrative Optimism component of Full-School Engagement.

Finally, some of the variance of each of the observed variables is unexplained by the other observed and latent variables in the model. This is called the *residual variance*. This variance has two components that cannot be distinguished by the model – other causes and purely random error. Virtually every model contains some degree of specification error; some of the variance in any observed variable can almost always be explained by variables that are either (a) not included in the model or (b) hypothesized as not affecting the observed variable. Ideally, the effects of these missing variables and paths are small. They constitute a part of the residual. This residual also contains measurement error. Even if all relevant variables and paths were included in a model, observed variables would often vary – partly because they are not measured with complete accuracy. In this model, the variance of 23 residuals (one for each variable, e1 to e23) is estimated. The model assumes that the covariance of these residuals with each other and with any latent variables in the model is zero. It also assumes that the residuals have a mean of zero. Twenty-three of the parameters of this structural equation model are the variances of residuals e1 to e23. Ideally, each residual variance will be small compared with the original variance of the observed variable it predicts. The ratio of

the residual variance of an observed, predicted variable to its overall variance (i.e., R^2) is an important measure of model fit. It answers the question: What percentage of variable *X*'s variance is explained by the model?

Variables in structural equation modeling are categorized as *endogenous* or *exogenous*. Exogenous variables have no arrows pointing to them; they are not predicted by anything in the model. Full-School Engagement is the only exogenous variable in this model. All other variables are endogenous. The observed endogenous variables ($y_{1\text{-}23}$) include residuals ($e_{1\text{-}23}$), as described above. The unobserved, or latent, endogenous variables ($\eta_{1\text{-}5}$) are associated with endogenous *disturbances* ($\zeta_{1\text{-}5}$), which function exactly like the residuals of observed variables.

Model 1, specified in Figure 2, is designed to be tested with a random quarter of the schools dataset, respecified as appropriate, and then retested with the entire schools dataset. To summarize the specification, a single second-order factor called Full-School Engagement is presumed to influence five first-order factors: Student Engagement, Student Resistance, Teacher Engagement, Parental Engagement, and Administrative Optimism. The primary hypothesis of this model, then, is that there is something important and measurable about a school that causes varying degrees of engagement in the educational process by all involved parties. This characteristic is denoted Full-School Engagement.

After specification, the **identification** of the model is investigated. In an identified model, a unique mathematical solution can be estimated for each of the estimated structural parameters. Bollen (1989) provides rules for the theoretical identification of a structural equation model. One necessary condition for the identification of a structural equation model is that each latent variable be provided a scale. In this study, each latent variable is scaled to

match the scale of one of its indicators. The scaling indicators have unstandardized[77] loadings fixed at 1. For this reason, they are indicated in Figure 2 with the numeral 1. The percentage of students absent on an average day provides a scale for Student Engagement. Reporting of problems with gang activities provides a scale for Student Resistance. Morale of teachers is the scaling indicator for Administrative Optimism. The percentage of teachers absent on a given day scales Teacher Engagement. Problems with parent involvement scale Parent Engagement. Finally, Teacher Engagement provides the scale for Full-School Engagement.

A second necessary condition for identification of a structural equation model is the *t-rule*. In order to be identified, a model must have at least one piece of information for each estimated parameter. With 23 observed variables, this model has 23 * 24 / 2 = 276 unique elements in the covariance matrix.[78] As described in Appendix A, the model estimates only 139 parameters, so the t-rule is amply satisfied. In fact, the model is overidentified.

We have seen that this model satisfies two necessary conditions for identification: its latent variables each have a scale and it has more than enough pieces of information to estimate its parameters. Even in combination, however, these necessary conditions do not suffice to guarantee that the model is identified. Bollen (1989) provides a series of sufficient conditions, but this model meets none of them. Therefore, empirical identification must suffice – the test of identification is the successful completion of a computational estimation algorithm such as that performed using Mplus and described in chapter 4.

---

[77] The standardized loadings of scaling indicators can differ from one, as shown, for example, in Table 10 and Figure 7..

[78] The formula for the number of unique elements in the covariance matrix for N variables is N(N+1)/2.

Description of CFA Variables and Estimation Method

Each of the 23 observed variables used in Model 1 comes from the 2003 NAEP grade 8 administrators survey. As described previously, each variable represents the opinion of a school administrator about the degree to which their school exemplifies a particular element of what is called Full-School Engagement in this study. Tables 5 to 9 show the distributions of administrator survey responses to the 23 questions these variables represent. All of the variables are categorical; most are skew. The analyses undertaken replace these categorical variables with estimated underlying continuous distributions. Weighted least squares is then used to obtain parameter **estimates**[79].

[79] The Mplus CATEGORICAL option and WLSMV estimator are used.

Table 5. Problems. Administrator survey responses to the question: "To what degree is each of the following a problem in your school?"[80]

Mplus Code	Item	Response			
		Not a Problem	Minor	Moderate	Serious
	Student Engagement				
Sabsprb	Student absenteeism	44%	42%	11%	2%
Strdprb	Student tardiness	28%	54%	16%	3%
	Student Resistance				
Fightprb	Physical conflicts among students	49%	44%	7%	1%
Raceprb	Racial or cultural conflicts	76%	22%	2%	0%
Gangprb	Gang activities	86%	12%	2%	0%
Smisbprb	Student misbehavior in class	20%	61%	17%	2%
Stftprb	Physical conflicts between students and teachers	91%	8%	1%	0%
Vandlprb	Vandalism	55%	41%	4%	0%
	Teacher Engagement				
Tabsprb	Teacher absenteeism	63%	30%	6%	1%
	Parent Engagement				
Pinvprb	Lack of parent involvement	32%	37%	23%	8%

[80] The percentages presented in Table 5 are from the full sample. The percentages in the 25% subsample are all within 2 percentage points of those from the full sample.

Table 6. Parent Involvement Percentages. Administrator survey responses to the question: "In your school, approximately what percentage of the parents do each of the following?"[81]

Mplus Code	Item	Response			
		0 - 25%	26 - 50%	51 -75%	76 - 100%
	Parent Engagement				
Currdec	Are involved in making school curriculum decisions	86%	10%	3%	2%
Opnhouse	Participate in open houses or back-to-school nights	11%	20%	34%	34%
Volnteer	Participate in volunteer programs	54%	25%	12%	10%
Ptconf	Participate in parent-teacher conferences	8%	18%	25%	49%
Pto	Participate in a parent-teacher organization	60%	18%	12%	10%

The **fit** of the model to the data is evaluated with global fit indicators CFI, TLI, and 1 - RMSEA. Parameters that differ significantly from zero in expected directions also support an argument that the model fits the data well. Finally, a well-fitting model should explain a good share of the variance of each of the endogenous variables. **Respecification** of the model is performed as needed to improve statistical fit and theoretical meaning. The model is re-estimated with a larger dataset. The final obtained results are **interpreted.** The interpretation of this CFA is primarily focused on the question it attempts to answer: *Can a single second-order latent variable called Full-School Engagement measure a constellation of factors representing administrative, parent, teacher, and student engagement in the academic mission of a school?*

[81] The percentages presented in Table 6 are from the full sample. The percentages in the 25% subsample are all within 3 percentage points of those from the full sample except for one. The category percentages for Ptconf in the 25% subsample are 7%, 19%, 30%, and 45%.

Table 7. Characterizations. Administrator survey responses to the question: "How would you characterize each of the following in your school?"[82]

Mplus Code	Item	Responses			
		Very Negative	Somewhat Negative	Somewhat Positive	Very Positive
		Student Engagement			
Sachatt	Students' attitudes toward academic achievement	0%	10%	59%	31%
		Student Resistance			
Propreg	Regard for school property	0%	7%	45%	47%
		Teacher Engagement			
Tmorale	Morale of teachers	0%	5%	39%	57%
Texpect	Teachers' expectations for student achievement	0%	3%	30%	68%
		Parent Engagement			
Parsupp	Parental support for student achievement	1%	7%	47%	45%

Table 8. Absentee Percentages. Administrator survey responses. "About what percentage of your…"[83].

Mplus Code	Item	Responses			
		0 - 2%	3 - 5%	6 - 10%	> 10%
		Student Engagement			
Sabspct	…students is absent on an average day?	38%	48%	13%	2%
		Teacher Engagement			
Tabspct	…teachers is absent on an average day?	76%	20%	4%	0%

[82] The percentages presented in Table 7 are from the full sample. The percentages in the 25% subsample are all within 3 percentage points of those from the full sample except for one set. The category percentages for Tmorale in the 25% subsample are 0%, 6%, 34%, and 60%.

[83] The percentages presented in Table 8 are from the full sample. The percentages in the 25% subsample are all within 4 percentage points of those from the full sample.

Table 9. Teacher retention percentages. Administrator survey responses to the question: "Of the full-time teachers who started in your school last year, what percentage left before the end of the school year?"[84]

Mplus Code	Responses						
	0%	1 - 2%	3 - 5%	6 - 10%	11 - 15%	16 - 20%	> 20%
Tquit	75%	19%	3%	2%	1%	0%	0%

Models 2, 3, and 4: Economic and Ethnic Composition Effects

Question 2 asks: *Do the economic or ethnic compositions of a school predict that school's mean grade 8 mathematics test scores, adjusted for the ethnicity and economic level of the individual students in that school?* This is a question about composition effects. It is answered by a sequence of three models. Model 2 is a simple regression model designed to replicate findings from many previous studies – student ethnicity and economic level predict test scores; neither fully explains the other. Model 3 is a baseline multi-level model that simply determines the percent of variance in student test scores that occurs between schools and the percent of variance that occurs within schools. Model 4 puts student ethnicity and economic level in the same model as school ethnic and economic composition, allowing the separation of effects required by Question 2, but not provided by hundreds of prior studies.

[84] The percentages presented in Table 9 are from the full sample. The percentages in the 25% subsample are all within 2 percentage points of those from the full sample.

Model 2: Baseline Regression

Model 2 tests the hypothesis that student ethnicity and economic level are associated with grade 8 NAEP mathematics test scores. A path diagram for this baseline regression is shown in Figure 3. Because ethnicity and economic level are both in the model, any effect found for ethnicity is controlled for economic level and any effect found for economic level is controlled for ethnicity. The SEM method used here is equivalent to a classic ordinary least squares linear regression with dummy variables.[85] The hypothesis is confirmed if (1) a large percentage of test score variance is predicted by the ethnicity and economic level variables, and (2) the estimated coefficients are statistically significant.

The model is first estimated with a random quarter of the combined dataset. The results are replicated with the remaining three-quarters of that dataset. Free-lunch status is used as a single variable to represent family income. It is expected to have a large and significant negative coefficient because it is a measure of the economic distance of the student from the tested middle-class school norm. Title I status is also expected to be negatively and significantly related to grade 8 mathematics test scores because students in schools with large numbers of low-income students receive Title I mathematics services on the basis of weak skills in the subject. Both the Title I and the free lunch results are expected even with the controls for ethnicity described below.

---

[85] In this case, the dummy variables for ethnicity are Hispanic, Black, Asian / Pacific Islander, American Indian or Alaskan Native, and Other. Each of these variables takes a value of 1 if the student is in that category and 0 otherwise. For White students, all of these indicators are 0. All ethnic parameter estimates therefore represent the distance from the White baseline. Title I status (1 for Title-I eligible students, 0 otherwise) is another dummy variable. Free-lunch status on the same 0-1 scale, but is not a true dummy variable because the value 0.5 is used for students with reduced-price lunch status. In a future study, the equal distance assumption implied by this coding will be tested using a pair of dummy variables for free-lunch and reduced-price-lunch status.

The other set of dummy variables provides information about the relative test performance of various U.S. ethnic groups, controlling for lunch and Title I status. For the purposes of NAEP, schools categorize students in one of five mutually exclusive ethnic categories: White, Hispanic, Black, Asian / Pacific Islander, and American Indian (including Alaska Native) (Rogers & Stoeckel, 2004, p. 30). Based on research presented in chapter 2, Hispanic, Black, and American Indian variables are expected to have significant negative coefficients, indicating that these ethnic groups perform more poorly on the grade 8 NAEP mathematics tests than White students, even when controlling for economic status. The coefficient for Asian / Pacific Islander is expected to be slightly greater than zero.

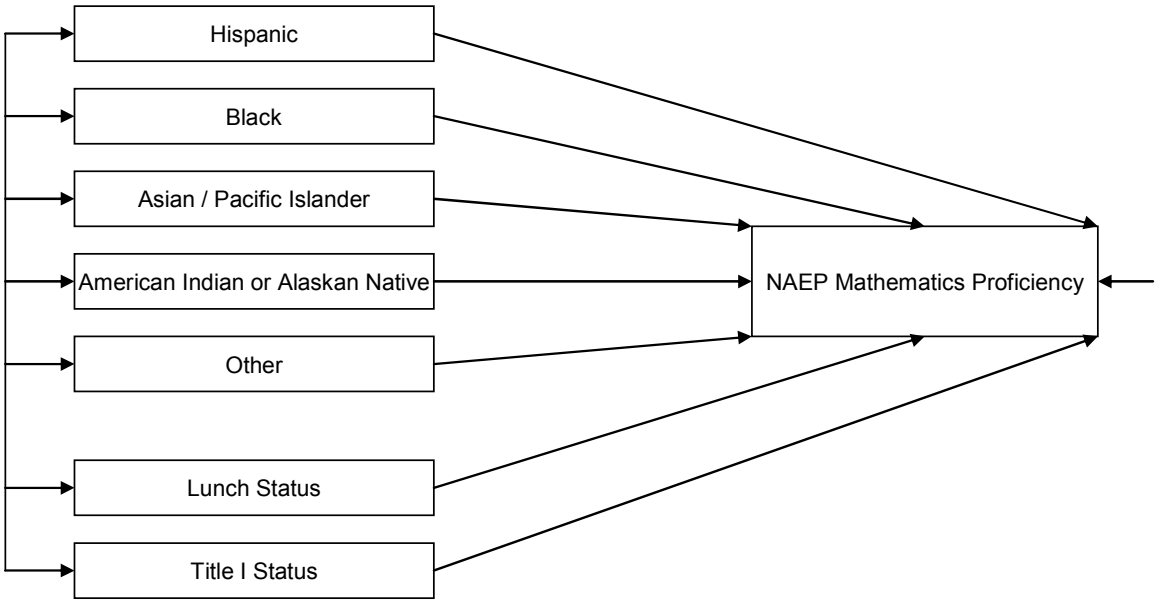
The outcome variable, NAEP mathematics proficiency, is a plausible value[86] for the student's mathematics scale score. NAEP scale scores are designed to measure how much students know and can do in mathematics, ranging on a scale from 0 to 500. IRT methods allow the use of the same scale for elementary, middle, and high school students. The grade 4 mean scale score is 235; the grade 8 mean scale score is 278 (Braswell, Dion, Daane, & Jin, 2005). Since the average student gains 43 points over these four years, one year's worth of growth is roughly equivalent to 10 or 11 points on this scale for grade 8 NAEP test-takers.

SEM is used to estimate this regression model. The results are identical to traditional ordinary least squares regression, but with the advantage of clarity. In SEM, all relationships intended by the model are included in the path diagram. The arrows pointing to NAEP Mathematics Proficiency from the economic and ethnic predictors represent the commonly interpreted regression parameters. The arrow pointing to the outcome variable with no visible source represents residual variance – the variance in the outcome variable not predicted by

[86] Plausible values are described in more detail in the final section of this chapter.

the model, some of which is measurement error and some of which is due to non-included variables. In traditional regression terminology, it is 1 - R^2. The set of arrows connecting the independent variables represents the standard regression assumption that independent variables are correlated.[87] Because each variable in a regression model is shown to be related to every other variable, the model is **just-identified**. Standard regression models are always just-identified. As described earlier, only the standard R^2 value[88] can be used to judge the overall fit of standard regression models; being just-identified, there are no excess data points that can be used for other tests of model fit.

Figure 3. Model 2. Baseline Regression.



Appendix A contains the equations and matrices associated with this model. The model is **estimated** using an Mplus maximum likelihood estimator with robust standard errors. Some

[87] The correlations and covariances of the independent variables are fixed to equal the correlations and covariances observed in the sample.

[88] R^2 is the square of the multiple correlation between an outcome varible (e.g., test score) and a set of predictor variables (e.g., ethnicity and economic level). It is also the percentage of variance in the outcome variable explained by the predictors. A high R^2 in a regression or an SEM is evidence of good model fit.

of the simpler models are also estimated with SPSS or SAS as a check. **Fit** is evaluated on the basis of the significance of the parameters and the amount of variance explained by the model (R^2). If necessary, the model is **respecified** and then retested with the remaining three-quarters of the dataset. The parameter estimates are **interpreted** to see how well they fit with prior research.

Model 3: Baseline Two-level Model

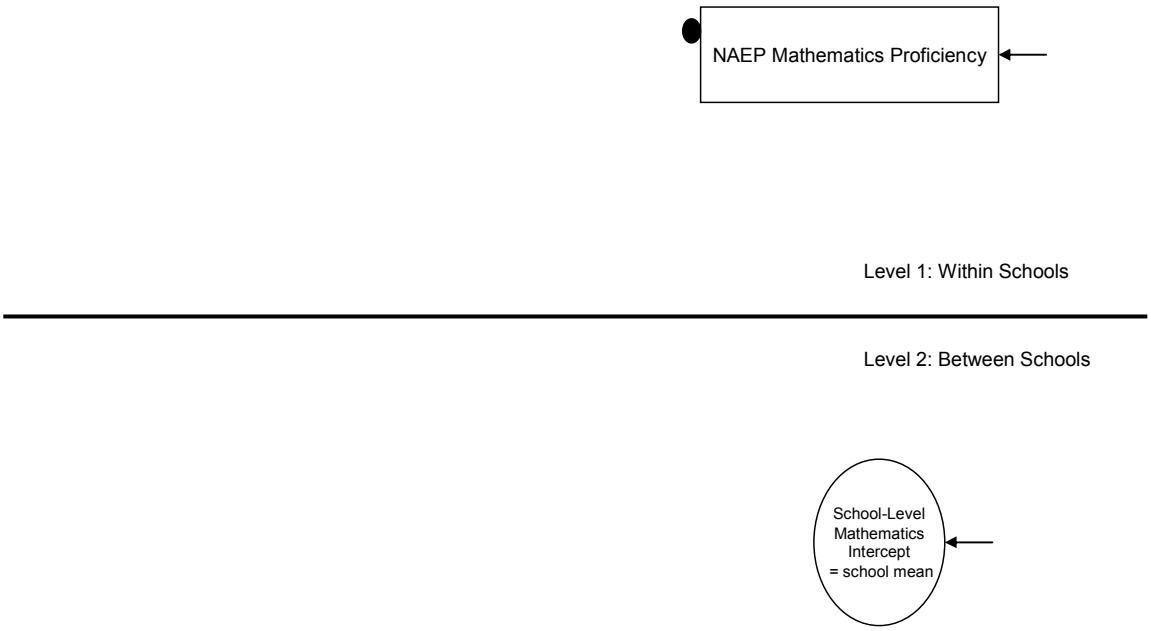
The third model (Figure 4) hypothesizes that variance in grade 8 NAEP mathematics test scores occurs both between and within schools. The amount of within-school variance is calculated by the level-one model. For each school in the sample, a test score mean and a variance around that mean are calculated. The arrow in the level-one model represents the average variance within a school[89] – the within-school variance. The solid oval represents the distribution of school means. The open oval in the level-two model represents the same distribution of school means. The arrow in the level-two model represents the variance in that distribution of school means – the between-schools variance.

Variance is expected both between schools (level 2) and within schools (level 1). This is generally judged by the value of the intra-class correlation (ICC, the ratio of the variance at level 2 to the total variance at the two levels). A low ICC value (e.g. below .10) is sometimes taken as a suggestion that the second level of a model is superfluous. Appendix A contains the equations, matrices, and assumptions associated with the third model.

[89] The within-school variance is a "precision-weighted average" (Raudenbush & Bryk, 2002, p. 40) of the variance within each school. Schools with larger sample sizes are given more weight because their estimates are more precise.

Theoretical **identification** of multi-level models has not been well studied. Empirical identification of Models 3, 4, and 5 must suffice. Model 3 is **estimated** with an Mplus maximum likelihood estimator. No global tests of model **fit** are appropriate for this model. No **respecification** will be needed. The **interpretation** is based purely on an estimate of the percentage of variance that occurs between schools, the ICC.

Figure 4. Model 3. Baseline Two-level Model.



Model 4: Composition Effects

The purpose of the fourth model is to perform a two-level separation of the relationships noted in Model 2. The simple regression of Model 2 is designed to document that student ethnicity and economic level are related to grade 8 NAEP mathematics test scores. Model 3 is designed to show that these test scores vary both between and within schools. Model 4 is a two-level economic and ethnic effects model. It is designed to test whether individual student ethnicity and economic level predict the grade 8 mathematics test scores of students within schools and whether school economic and ethnic compositions predict grade 8 adjusted

school mean NAEP mathematics test scores between schools. As with Model 2, within-school and between-school effects are separated as they are analyzed in the same model.

This analysis is performed on the same random quarter of the combined dataset used with Models 2 and 3, then retested on the remaining three quarters of the data. Six new variables are added. School-record-based Percent Asian, Percent Black, Percent Hispanic, and Percent Indian variables are placed on a decimal scale ranging from zero to one. Percent free or reduced-price lunch and percent Title I are each recoded from a 1-8 scale to a 0-1 scale by replacing each categorical value with the midpoint of the range it represents. For example, administrators are asked: "During this school year, about what percentage of students in your school was eligible to receive a free or reduced-price lunch through the National School Lunch Program?" A response of (d) corresponds to "11 – 25%" and is coded as "4" in the NAEP dataset. In this analysis, that "4" is replaced with .18, the midpoint of the 11-25% range. These transformations make each of the six composition variables effectively continuous on a 0-1 scale, with 0 representing no students in the given category and 1 representing 100%.

The variables representing within-school ethnicity and economic level are also transformed for this model. They are all grand mean centered for ease of interpretation, as described below. With this recoding, there are two possible values for each of the individual ethnicity variables. About 15% of students are Black. They receive a value of .85. About 85% of students are not Black. They receive a value of -.15. The mean of these values across all records in the dataset is $(.85)(-.15) + (.15)(.85) = 0$. The dataset mean of any grand mean centered variable is zero. The hypothetical average student in this modeling strategy, with a

value of 0 for the variable Black, is 15% Black, even though no student in the sample is listed with such a value.

Level one of a two-level model can be thought of as a weighted mean of a set of regressions (one regression per school). Each regression has an intercept. In this study, the intercept is the predicted mathematics test score of a student at the school with values of zero for all six independent variables. Since the variables are all grand-mean centered, this is the predicted score of a hypothetical student at that school who is nationally average both ethnically[90] and economically.[91] This is the grade 8 adjusted school mean NAEP mathematics test score. For a low-income school, this intercept is expected to be higher than the actual school mean because the average student in the nation is more likely to be non-poor than the average student at that school, and non-poor students tend to score higher than poor students at almost all schools. By focusing on this intercept rather than true school mean, grand-mean centering controls for the aggregated ethnicity and economic levels of the students in each school, allowing a focus on true composition effects.

It is, of course, not a foregone conclusion that the specified individual and composition effects all exist. It is a hypothesis that is tested by this model. The expectation is that some of the variables at both levels will show significant relationships with grade 8 NAEP mathematics test scores. Like Models 2 and 3, this model is **just-identified**. Therefore, there will be no meaningful test of overall **fit**, but, along with some significant relationships,

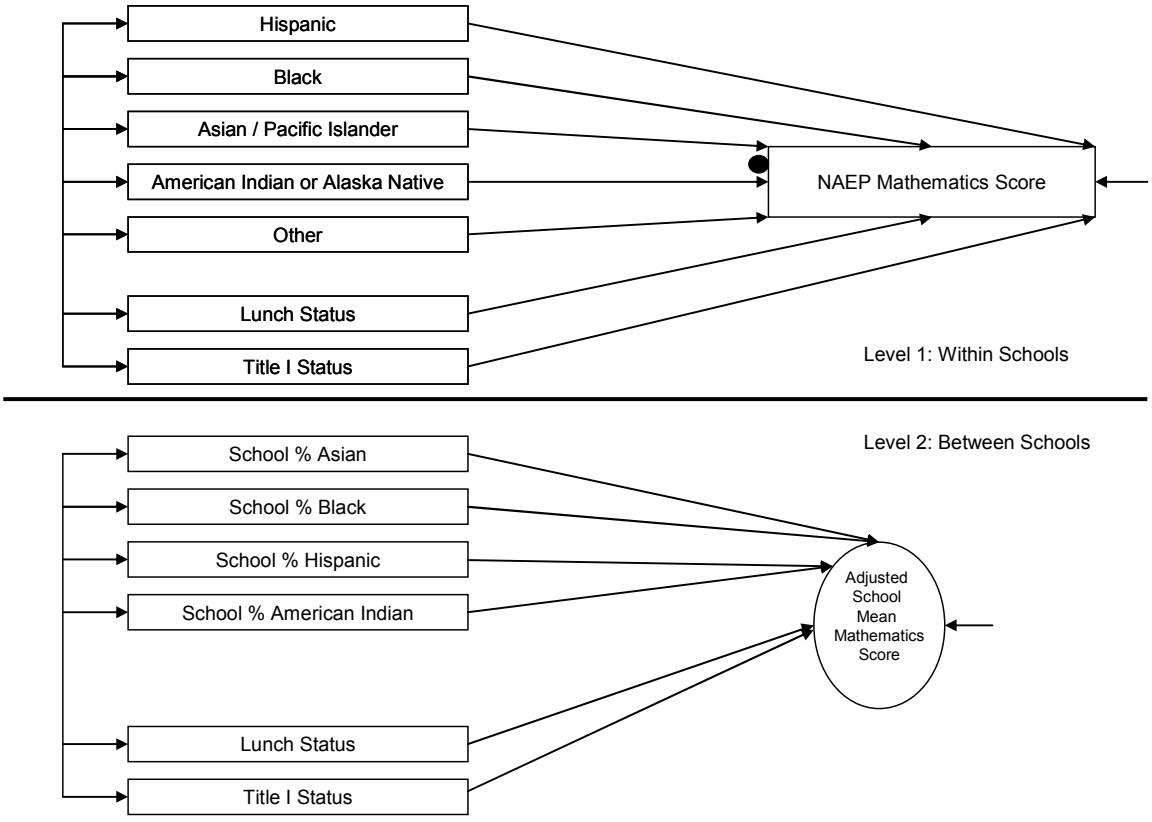
---

[90] The national average student would be about 2% Indian, 10% Hispanic, 15% Black, 3% Asian, and 70% White and Other, according to the estimations of the three-quarter replication dataset.

[91] The national average student would have a score of .39 on the free lunch scale, placing him slightly above the reduced-price lunch income level. The student would be 24% Title I status. It is important to emphasize again that these "national average" students are hypothetical; there is not a single student in the dataset that is national average because each of the variables involved has only two values. One value is higher than the national average; the other is one point below the high value and lower than the national average.

reasonably large R^2 values are expected. It is **estimated** with an Mplus maximum likelihood algorithm. Because of the theoretical importance of each of the predictor variables, no **respecification** will be attempted. It is estimated twice, once with a random quarter of the combined dataset and then again with the remaining three-quarters of that dataset. The **interpretation** of the model will be substantive and interesting because the literature on composition effects is weak, particularly regarding smaller ethnic groups. The parameter estimates allow a much- needed separation of within-school and between-school effects and, simultaneously, of ethnic and economic effects.

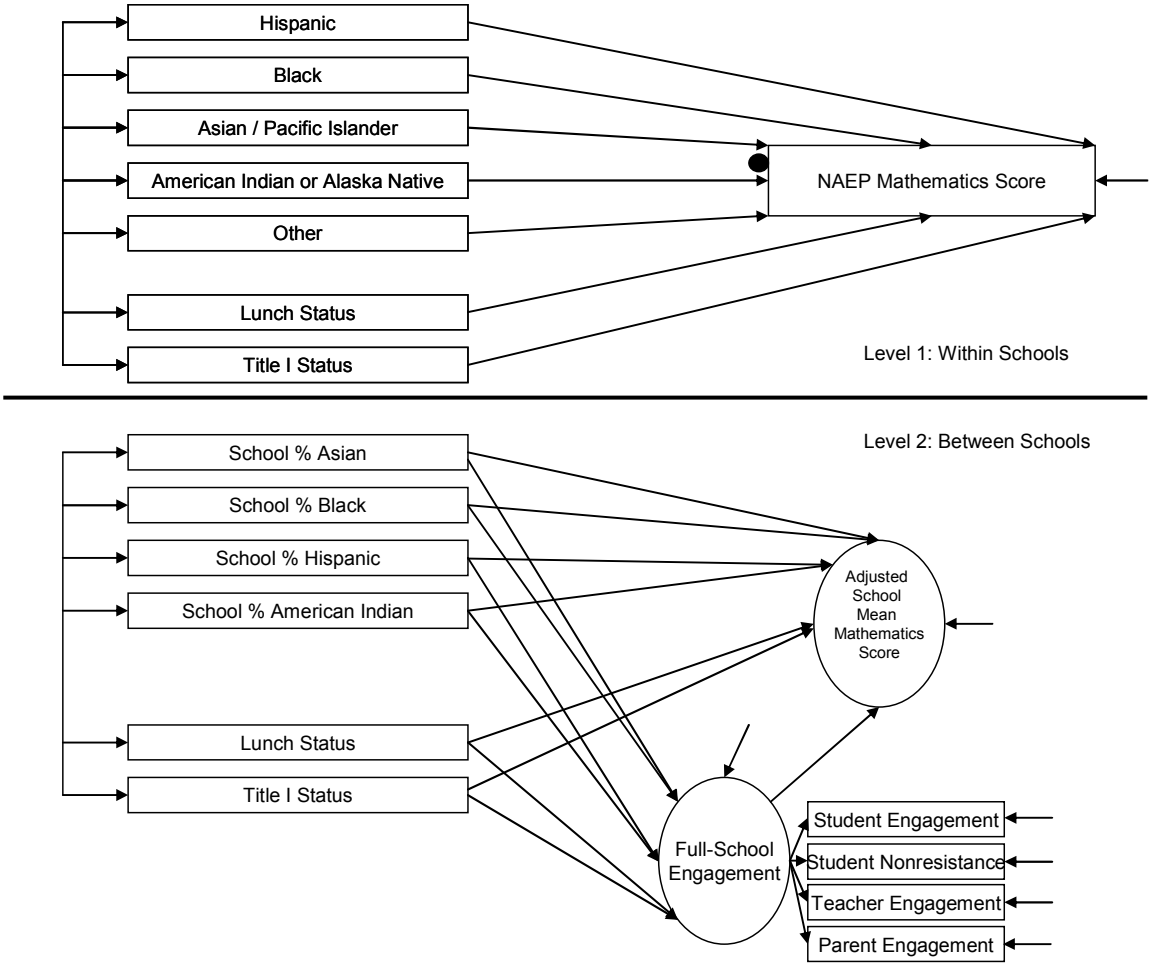
Figure 5. Model 4. Composition Effects Model



The darkened oval in the level one model, as well as the open oval in the level two model, represents the distribution of intercepts between schools. Therefore, in this grand-mean-centered model, the coefficients of the between-school predictor variables represent pure

composition effects – the effects of school economic and ethnic compositions on the school grade 8 NAEP mathematics score intercept, where the intercept is a mean that is adjusted for the effects of individual student ethnicity and economic level (for a more complete explication of this topic, see Raudenbush & Bryk, 2002, pp., pp. 139-141). The model contains six such composition effects. Appendix A contains the equations, matrices, and assumptions associated with this model.

Figure 6. Model 5. Mediation Model



Model 5: Mediation Model

By adding Full-School Engagement to Model 4, the fifth and final model can address Question 3: *What are the relationships that exist among the economic and ethnic compositions of a school, Full-School Engagement, and adjusted school mean grade 8 mathematics test scores?* Viewed in combination with Model 4, Model 5 can also answer Question 4: *Does Full-School Engagement mediate any of the composition effects identified in Question 2?*

The hypothesis of the model is that the addition of Full-School Engagement to Model 4 will decrease the direct economic and ethnic composition effects on grade 8 NAEP mathematics test scores, as these effects are partially replaced by significant indirect effects through the Full-School Engagement variable. To support this hypothesis, the model fit should be good; Full-School Engagement should be predictive of adjusted school mean grade 8 NAEP mathematics test score and well-predicted by composition variables (as measured by R^2 values and significance of regression coefficients). If all hypotheses are supported by the data, evidence will be provided that differential Full-School Engagement in segregated schools is part of the explanation for the mathematics test score gaps that characterize our educational system and nation.

The addition of the Full-School Engagement variable as a mediator between each of the school composition variables and the outcome variable is the only **specification** difference between Model 5 and Model 4. Ideally, the Full-School Engagement latent variable would be estimated within the context of the full model presented in Figure 6. This may be possible with more computing power, but this study requires a compromise. The four primary first-order latent variables from Model 1 (Student Engagement, Student Resistance, Teacher

Engagement, and Parent Engagement) are each replaced with a sum-score variable. The values of the indicator variables for each are summed[92] to create these new observed variables. Because it shares indicators with the other variables, Administrative Optimism is removed from the model.

This model satisfies the t-rule for identification. The difference between pieces of information available from the covariance matrix and the number of parameters to be estimated is 23. The model is therefore overidentified. It satisfies a necessary, but not sufficient, condition for identification. **Identification** is empirically tested by the process of estimation. An Mplus maximum likelihood **estimator** is used. As an overidentified model, global tests of model **fit** are available and are examined. Additionally, parameters and R^2 values are examined to consider model fit. **Respecification** is considered; the model is retested with a second three-quarter sample and **interpreted**.

This most complete model is subjected to the most detailed interpretation. Question 3 is answered: *What are the relationships that exist among the economic and ethnic compositions of a school, Full-School Engagement, and adjusted school mean grade 8 mathematics test scores*? The significance levels and actual levels of the paths in the model are interpreted. If, as hypothesized, Full-School Engagement partially mediates the composition effects of school ethnic or economic composition found in Model 4, then the following should occur:

a) The path between Full-School Engagement and Adjusted School Mean Grade 8 NAEP Mathematics Score should be significantly greater than zero.

---

[92] Prior to summing, the signs of some of the 23 indicator variables are reversed to ensure that the variables represent positive engagement. In particular, all indicators of problems are reversed, as are indicators of tardiness, absence, and departure from the job. The Student Resistance variable becomes Student Nonresistance in the process, orienting all four latent variables in a positive direction as measures of Full-School Engagement.

b) The path from the composition variable in question to Full-School Engagement should differ significantly from zero.

c) The path from the composition variable in question to Adjusted School Mean Grade 8 NAEP Mathematics Score should be weaker in Model 5 than in Model 4, as part of the variable's influence is explained by Full-School Engagement.

If such results are found for some of the school composition variables, then the central hypothesis of this study will be supported. Full-School Engagement will partly explain why schools with more students from low income levels or certain ethnic groups tend to have lower scores on mathematics tests. Appendix A contains the equations, matrices, and assumptions associated with this model.

Section 3. Technical Issues and Software Choice

The analysis of these data requires effective handling of numerous technical issues. First, NAEP data are stratified, not a simple random sample. This requires the use of **weights**. Second, there is significant clustering in the dataset. Clustering and non-random sampling can affect variances. NAEP developers recommend **jackknife variance estimation**. Third, the outcome variables are **plausible values**, not accurate point estimates of student proficiency. Fourth, many of the **variables are categorical** rather than continuous. Fifth, some of the variables include a significant amount of **missing data**. Finally, the **data are cross-sectional**, not longitudinal. *Mplus* software is chosen because of its ability to handle the majority of these problems while estimating a two-level structural equation model.

Weights

NAEP provides school-level and student-level weighting variables that adjust for probability of selection. The student-level weights are used for the student-level basic regression (Model 2). School-level weights are clearly appropriate for the school-level confirmatory factor analysis (Model 1); they are also recommended by *Mplus* staff for two-level models (personal correspondence, L. Muthén). *Mplus* software allows for the use of weighting variables.

Jackknife Variance Estimation

NAEP recommends calculation of jackknife variance estimates (Efron, 1979; Little & Rubin, 2002; Rogers & Stoeckel, 2004) because of clustering (students clustered within schools) and other adjustments to NAEP data. Analysts are instructed to run the model with standard weights, then 62 more times, each using one of the 62 sets of replicate weights provided with the dataset. The sample variance of each parameter is then estimated as the sum of the squared differences of its 62 estimates from the baseline run. The time-consuming jackknife is used for Models 1 and 5. Analyses reported in chapter four suggest that two-level analysis largely obviates the need for jackknife variance estimation.

Plausible Values

NAEP's plausible values are fully described in a 1992 article from the Educational Testing Service (ETS) (Mislevy, Beaton, Kaplan, & Sheehan, 1992). In order to keep tests short for individual test-takers while covering all important topics, test-takers are instructed to respond

to only a small portion of the mathematics items, making accurate estimation of their individual proficiency impossible. Missing-data imputation techniques (Little & Rubin, 2002) are combined with IRT proficiency estimation methods to generate a set of five plausible ability values for each student. These variables are named mrpcm1, mrpcm2, mprcm3, mrpcm4, and mrpcm5. Whenever a test score is used as a variable in a NAEP model, that model should be run five times; once for each plausible value. The parameter estimates should be averaged and the variability due to use of plausible values[93] should be added to estimates of sampling variability, despite increasing standard error and decreasing the likelihood of significant results. Plausible values are used in this manner for the final model only. For Models 2, 3, and 4, a single plausible value is used. Analyses reported in chapter four suggest that this approach does not affect any conclusions of the study.

Categorical Variables

In traditional SEM, observed dependent variables are expected to be continuous and normally distributed[94]. Many of the observed variables in this model do not meet this assumption. They are ordered categorical variables with four to seven categories (see Appendix B). Modeling them without accounting for their categorical nature would yield incorrect results (Bollen, 1989, pp. 438-439). One solution to this problem is to replace categorical observed variables (y) with latent, normal, continuous variables (y^*) by defining a continuous distribution and thresholds (τ) in the distribution at which each category is

[93] The variability due to the use of five plausible values is 1 1/5 times the observed variability of those plausible values (Mislevy et al., 1992; Rogers & Stoeckel, 2004).

[94] A slightly less restrictive assumption is that the variables have no excess multivariate kurtosis.

observed (B. O. Muthén, 1998-2004, pp. 1-5). This adds a threshold model to the measurement model (Bollen, 1989, p. 441). Traditional maximum likelihood estimation generates inaccurate significance tests and fitting functions in this case. Weighted least squares estimation performs better (Bollen, 1989, p. 443). *Mplus* is one of the few structural equation modeling packages capable of modeling categorical dependent variables appropriately. The Mplus CATEGORICAL option performs these calculations. The WLSMV[95] estimation choice has greater computational efficiency than WLS because it uses only the diagonal of the weight matrix in computations, thereby removing the computationally intensive need to invert a large matrix. Adjustments to mean and variance estimates are made to account for this simplification. WLS estimation of large models such as this requires large samples. Even a quarter of the NAEP school sample is easily large enough.

Missing Data

The variables included in this study contain between 1 and 15 percent missing data. Ideally, multiple imputation or maximum likelihood would be used to fill these holes or estimate around them (Allison, 2003; Little & Rubin, 2002; Peugh & Enders, 2004). Unfortunately, NAEP's scale-score plausible values are created via multiple imputation and presented to the researcher as a *fait accompli*. Re-creation and modification of this imputation process is beyond the scope of this study because the IRT models used to create the scale scores would have to be recreated simultaneously. Maximum likelihood missing-data methods in combination with categorical data estimation were too computationally

[95] Other writers call this estimation technique diagonally weighted least squares or DWLS.

intensive for this model with *Mplus 3.13*. Therefore, the blunter instruments of pairwise and listwise deletion are used. The amount of data lost in this process is reported by model in chapter four.

Cross-Sectional Data

Finally, NAEP provides only cross-sectional data. Most important for studies of academic achievement, there is no pre-test. These facts require that any causal inferences drawn from this study be tentative. A particular threat to this analysis is the possibility of reverse causation. Perhaps higher mathematics test scores lead to higher Full-School Engagement instead of the opposite. This may happen if students with less academic ability are assigned to schools with lower Full-School Engagement or if school communities react to low ability by disengaging. The literature reviewed suggests the proposed causal model, but these possibilities are also plausible. The reverse-arrow alternate model will be tested and compared to Model 5. In general, the problems with causal inference are ameliorated, but by no means removed, by the modeling of multiple paths in SEM. Nevertheless, acceptance of the models in this study can not be taken to imply that these are the only possible ways to model the data. In fact, there are always many ways that the same data can be effectively modeled. Hopefully, the literature reviewed, along with my experience in schools, has allowed me to choose a defensible model, but I do not expect the results presented in chapter four to close any debates. I will be happy if they open new ones.

CHAPTER FOUR - RESULTS

Overview

This study uses a series of five models to investigate the role of Full-School Engagement in relation to ethnic and economic test score gaps in the U.S. All of the models use grade 8 NAEP mathematics data from the 2003 assessment. As described at the beginning of Chapter 3, the models are (1) a confirmatory factor analysis of Full-School Engagement, (2) a baseline student-level regression model of the relationships of student ethnicity and economic level with grade 8 mathematics test scores, (3) a baseline two-level model of the mathematics test scores of grade 8 students clustered within schools, (4) a composition effects model of the relationships of school ethnic and economic composition with grade 8 school mean mathematics test scores, adjusted for the ethnicity and economic levels of individual students, and (5) a mediation model in which Full-School Engagement explains a part of the composition effects found in Model 4. Descriptions of the samples and distributions of key variables are provided in Chapter 3, as is a discussion of the identification of each specified model.

Model 1: Measuring Full-School Engagement

Model 1 is originally specified by the path model in Figure 2 and the equations and matrices in Appendix A. The purpose of the model is to answer Question 1: *Can a single second-order latent variable called Full-School Engagement measure a constellation of factors representing administrative, parent, teacher, and student engagement in the academic mission of a school?* The original model of Full-School Engagement (fullcfa6b2) is tested

and estimated with a random quarter of the schools dataset. A re-specified and improved model (fullcfa7b1) is then tested and estimated with the same data. A new one-quarter sample is drawn for a replication of the improved model (baseline25), and finally, the full dataset is used for an estimation that includes the most data and also the most accurate and laborious estimation of standard errors. Each of these four estimations is described in this section.

The original model, named fullcfa6b2, is simultaneously tested and estimated with Mplus software on a random quarter of the NAEP Grade 8 schools dataset. A series of tests is performed to verify that the model is identified (i.e., estimable) and that it matches the data well. At the same time, estimates of the parameters specified in the model are generated. The results are presented in Figure 7 and, equivalently, in column 1 of Table 10.

The first section of Table 10 provides basic information about the model. The estimator used by the Mplus program to test and estimate the model is WLSMV.[96] Missing data are pairwise deleted.[97] Overall, the dataset contains 1521 records, representing 1521 schools. The most accurate method of variance and standard deviation estimation for the NAEP dataset is a computationally intensive method called jackknife. This method is not used for the two preliminary models.

---

[96] WLSMV generates weighted least squares parameter estimates (thus WLS). It uses a diagonal weight matrix to speed computation. A full-weight matrix is used to calculate standard errors and a chi-square statistic that is mean- and variance-adjusted (thus the MV).

[97] The estimation of this model is based on the variances and covariances of the 23 observed variables with each other. With *pairwise deletion*, each variance is calculated based upon all records that contain data for the variable in question; each covariance is calculated based upon all records that contain data for the pair of variables in question. The *N* reported by Mplus in the case of pairwise deletion is the number of records in the entire dataset (e.g. 1521 in the first two columns of Table 10). An alternative used with other models in this section is *listwise deletion*, in which only records missing no data at all are included in the analysis. In these cases, the reported *N* is only the records used (e.g. 1430 in the third column of Table 10).

Figure 7. Results of initial Full-School Engagement Confirmatory Factor Analysis

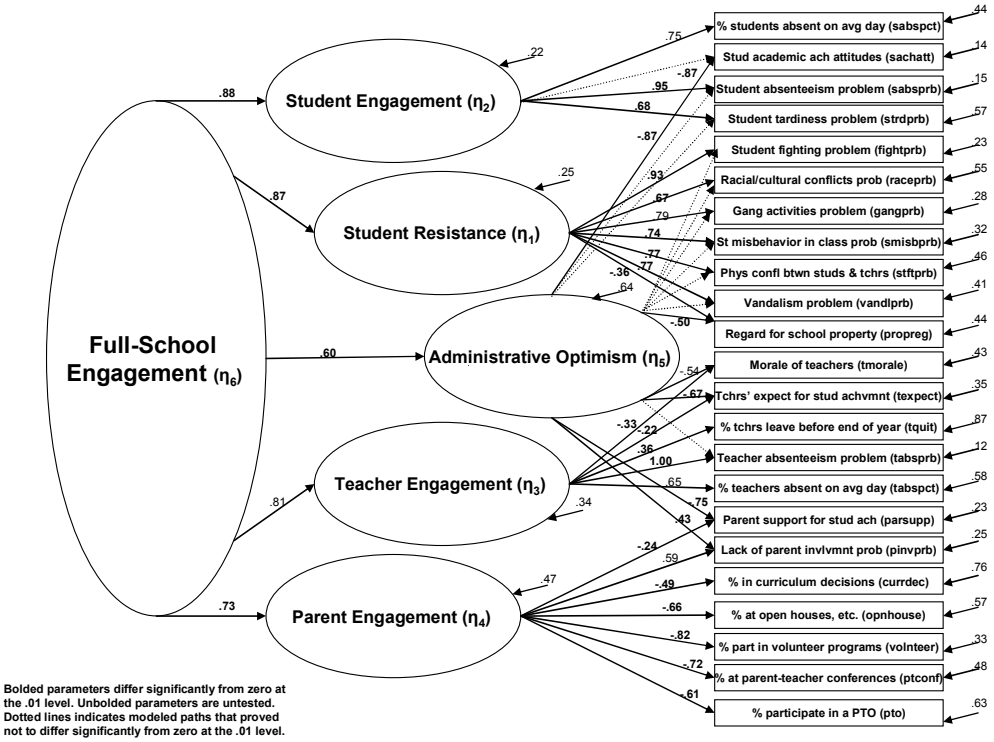


Table 10. Model 1. Confirmatory Factor Analysis of Full-School Engagement

Label	Model Name			
	fullcfa6b2	**fullcfa7b1**	**baseline25**	**baseline100**
Estimator	WLSMV	WLSMV	WLSMV	WLSMV
Missing	pairwise	pairwise	listwise	listwise
N	1521	1521	1430	5687
Jackknife	no	no	yes	yes
Overall Fit Indices				
Chi-Sq	150.4	144.6	135.0	526.5
df	44	45	44	72
p-value	0.000	0.000	0.000	0.000
CFI	0.961	0.963	0.964	0.960
TLI	0.973	0.975	0.976	0.978
1-RMSEA	0.960	0.962	0.962	0.967
SRMR			0.068	0.055
Measures of Student Engagement (Seng)				
~~***sabspct***~~	~~0.75~~	~~0.75~~	~~0.74~~	~~0.67~~
sachatt	*-0.10*			
sabsprb	**0.95**	**0.91**	**0.92**	**0.90**
strdprb	**0.68**	**0.65**	**0.65**	**0.75**

Label	Model Name			
	fullcfa6b2	fullcfa7b1	baseline25	baseline100
Measures of Student Resistance (Sres)				
fightprb	**0.93**	**0.86**	**0.86**	**0.85**
raceprb	**0.67**	**0.69**	**0.69**	**0.73**
~~***gangprb***~~	~~0.79~~	~~0.76~~	~~0.76~~	~~0.74~~
smisbprb	**0.74**	**0.70**	**0.71**	**0.73**
stftprb	**0.77**	**0.74**	**0.74**	**0.77**
vandlprb	**0.77**	**0.78**	**0.79**	**0.74**
propreg	**-0.36**	**-0.32**	**-0.34**	**-0.26**
Measures of Teacher Engagement (Teng)				
tmorale	**-0.33**	**-0.29**	**-0.29**	**-0.27**
texpect	**-0.22**	-0.18	*-0.14*	**-0.19**
tquit	**0.36**	**0.37**	**0.36**	0.36
tabsprb	**1.00**	**0.89**	**0.91**	**0.91**
~~***tabspct***~~	~~0.65~~	~~0.67~~	~~0.67~~	~~0.65~~
Measures of Parent Engagement (Peng)				
parsupp	**-0.24**	**-0.21**	**-0.22**	**-0.25**
~~***pinvprb***~~	~~0.59~~	~~0.54~~	~~0.53~~	~~0.57~~
currdec	**-0.49**	**-0.50**	**-0.50**	-0.38
opnhouse	**-0.66**	**-0.66**	**-0.66**	**-0.63**
volnteer	**-0.82**	**-0.83**	**-0.83**	**-0.77**
ptconf	**-0.72**	**-0.73**	**-0.74**	**-0.78**
pto	**-0.61**	**-0.62**	**-0.61**	**-0.68**
Measures of Administrative Optimism (Admopt)				
sachatt	**-0.87**	**-0.94**	**-0.95**	**-0.95**
strdprb	*-0.04*			
fightprb	-0.12			
raceprb	*0.01*			
gangprb	*0.11*	0.15	0.16	*0.10*
smisbprb	0.15	**0.20**	**0.19**	0.16
stftprb	*-0.06*			
vandlprb	*-0.01*			
propreg	**-0.50**	**-0.52**	**-0.53**	**-0.52**
~~***tmorale***~~	~~-0.54~~	~~-0.56~~	~~-0.56~~	~~-0.55~~
texpect	**-0.67**	**-0.70**	**-0.74**	**-0.69**
tabsprb	*-0.16*			
parsupp	**-0.75**	**-0.76**	**-0.76**	**-0.75**
pinvprb	**0.43**	**0.47**	**0.48**	**0.48**
Measures of Full-School Engagement (FSE)				
seng	**0.88**	**0.89**	**0.89**	**0.86**
sres	**0.87**	**0.86**	**0.85**	**0.87**
~~***teng***~~	~~0.81~~	~~0.80~~	~~0.79~~	~~0.76~~
peng	**0.73**	**0.72**	**0.72**	**0.65**
admopt	**0.60**	**0.61**	**0.61**	**0.62**

Label	Model Name			
	fullcfa6b2	fullcfa7b1	baseline25	baseline100
Percent of variance explained by model (R^2)				
sabspct	0.56	0.57	0.55	0.45
sachatt	0.86	0.89	0.89	0.90
sabsprb	0.85	0.84	0.84	0.81
strdprb	0.43	0.42	0.43	0.56
fightprb	0.77	0.74	0.74	0.72
raceprb	0.45	0.47	0.47	0.53
gangprb	0.72	0.73	0.74	0.65
smisbprb	0.68	0.67	0.68	0.68
stftprb	0.56	0.55	0.55	0.59
vandlprb	0.59	0.61	0.62	0.54
propreg	0.56	0.56	0.57	0.48
tmorale	0.57	0.56	0.56	0.52
texpect	0.65	0.63	0.66	0.63
tquit	0.13	0.14	0.13	0.13
tabsprb	0.88	0.78	0.72	0.82
tabspct	0.42	0.45	0.45	0.42
parsupp	0.77	0.75	0.76	0.77
pinvprb	0.75	0.74	0.74	0.77
currdec	0.24	0.25	0.25	0.14
opnhouse	0.43	0.44	0.43	0.40
volnteer	0.67	0.68	0.69	0.60
ptconf	0.52	0.54	0.55	0.60
pto	0.37	0.38	0.37	0.46
seng	0.78	0.78	0.80	0.74
sres	0.75	0.73	0.73	0.75
teng	0.66	0.64	0.62	0.58
peng	0.53	0.51	0.51	0.42
admopt	0.36	0.37	0.37	0.39

Note. All parameter estimates are standardized. Bold font represents parameter estimates that differ significantly from zero at the .01 level. Standard font represents estimates that differ significantly from zero at the .05 level. Italics are used for parameter estimates that do not differ significantly from zero. Strikethrough font is used to represent parameter estimates that are not significance tested; these parameters are generally used to scale the factors they measure.

The second section of Table 10, labeled *Overall Fit Indices*, provides a series of measures of the overall fit of the proposed model to the data. The most basic measure of the overall fit of a structural equation model to a dataset is a chi-square test. A high value suggests that the actual covariance matrix differs from the covariance matrix derived computationally from

model assumptions. Fullcfa6b2 produces a chi-square value of 150.4, with 44 degrees of freedom. This is highly significant (p=0.000). Some structural equation modelers consider such a value to be a hard disconfirmation of the model, but most recognize that such highly significant values are almost always obtained when working with large samples. They are taken to imply that the model is imperfect, but not that it is untenable.

Other fit indices, not so strongly influenced by sample size, have been developed. Each index has its own strengths and weaknesses. Bollen (1989, p. 281) recommends several of them. Mplus provides CFI, TLI, and 1 – RMSEA. Each of these is considered strong if it is greater than .95. These alternative fit indices support the proposed model (CFI = .961, TLI = .973, 1 – RMSEA = .960).

Strong measures of overall fit are not sufficient support for a model. Model parameters should also have signs and magnitudes that are reasonable in the context of theory and prior research. (Refer to Figure 7 and the first column of Table 10 as this review of model parameters is described.) With one exception, the significance tested[98] measures of the primary factors (Student Engagement, Student Resistance, Teacher Engagement, and Parent Engagement) in fullcfa6b2 differ significantly from zero at the .01 level[99] in consistent directions,[100] with standardized loadings ranging in absolute value from .22 to 1.00[101]. This

[98] As described in Chapter 3, one measure of each construct is not significance tested because it is used to scale the indicator and is thus not free to vary.

[99] In figures and tables in this document, bold font is used to indicate parameters whose values differ significantly from zero at the .01 level.

[100] The directions of the coefficients in this model make clear that high values for student, teacher, and parent engagement actually represent disengagement. With this interpretation, Full-School Engagement is, in fact, Full-School Disengagement. This wholesale reversal of signs is an artifact of the specification and has no negative impact on the interpretation of the results.

suggests that appropriate measures were chosen for each factor. In Figure 2, each of these paths is labeled with its standardized loading. The modeling of Student Engagement is an example.

The parameters that link Student Engagement to its indicators are called *factor loadings* in factor analysis tradition, but they can also be thought of as standardized regression coefficients or beta weights. The .75 coefficient of Student Engagement measured by the percentage of students absent on an average day (sabspct) says that an increase of 1 standard deviation in Student (dis)Engagement is associated with an increase of .75 standard deviations in the reported number of students absent on an average day.

Student Engagement explains $.75^2 = 56\%$ of the variance in the sabspct variable. This "percentage of variance explained by the model" is also known as R^2 and is found in the final section of Table 10. As would be expected, some of the variance of each observed variable remains unexplained by the model. As described in Chapter 3, these *residuals* (e1 – e23 in Figure 2) represent measurement error and the effects of variables and paths not included in the model. The residual for sabspct is named *e1*. The standardized variance of the residual, $1 - R^2$, is .44 in this case and is found in the far right column of Figure 7. No significance test is possible for the loading of sabspct on Student Engagement because sabspct is used to scale Student Engagement. For this reason, the parameter estimate is written in strikethrough font in Table 10 and is not bolded in Figure 7.

Table 10 shows the estimated loading of *student academic achievement attitudes* (sachatt) in italics; Figure 7 shows the causal arrow connecting *student academic achievement*

---

[101] When a single variable is used to measure more than one construct, its standardized loading can be 1.00 or higher. This is the case with the loading of teacher absenteeism problem (tabsprb) on teacher engagement. (Jöreskog, 1999)

attitudes to Student Engagement as dotted with no loading. The italics in the table and the dotted arrow in the figure show that the loading (-.10) does not differ significantly from 0 at the .01 level. The reason for this weak loading may be that administrative reports of student academic achievement attitudes are very subjective. Administrative Optimism, with a significant coefficient of -.87, appears to be much more strongly predictive of an administrator's characterization of students' attitudes toward academic achievement than is Student Engagement. The specification weakness represented by this loading is addressed in the respecification that leads to the fullcfa7b1 model.

The third indicator, *student absenteeism problem* (sabsprb) has a significant loading of .95, while the fourth, *student tardiness problem* (strdprb), has a significant loading of .68. The next three sections of the first column of Table 10 show that all the testable measures of *Student Resistance*, *Teacher Engagement*, and *Parent Engagement* are significantly related to their associated constructs. Furthermore, as seen in the last section of the table, the model explains between 13 percent (tquit) and 88 percent (tabsprb) of the variance observed for each variable. The final column of Figure 7 shows that this leaves between 12 percent (tabsprb) and 87 percent (tquit) of variance in each variable unexplained.

Administrative Optimism is not measured quite as well as the four primary constructs in this first model. It is clear from a comparison of loading significance that most of the observed variables measure their primary construct more than they measure *Administrative Optimism.* Only five of the tested observed variables are significantly related to *Administrative Optimism.* The variables sachatt (-.87), propreg (-.50), texpect (-.67), parsupp (-.75), and pinvprb (.43) are the significant indicators of Administrative Optimism. The variable tmorale, the scaling indicator, is not tested, but 57% of its variance is explained by

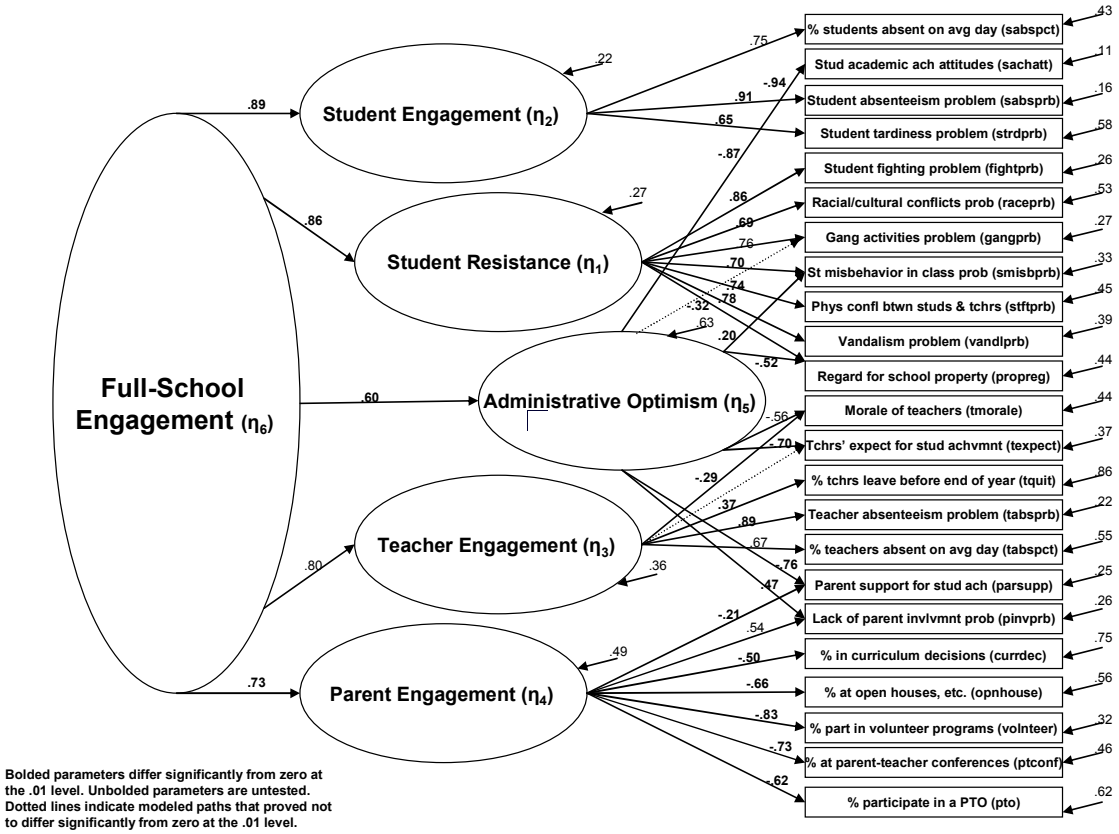
the combination of *Administrative Optimism* and *Teacher Engagement*. The directions of the loadings are consistent.[102] As indicated by dotted lines in the figure and italics in the table, eight indicators of Administrative Optimism do not differ significantly from zero at the .01 level. These eight (student tardiness problem, student fighting problem, racial/cultural conflicts problem, gang activities problem, student misbehavior problem, physical conflict between students and teachers problem, vandalism problem, and teacher absenteeism problem) seem to be, in general, more objective measures and thus less affected by the optimism or pessimism of the administrator completing the survey. The five first-order factors all load significantly on Full-School Engagement (FSE). The range is .60 (admopt) to .88 (seng). Overall, model fullcfa6b2, based on the initial specification, seems promising but not perfect – a strong candidate for respecification.

A respecified, more parsimonious model, fullcfa7b1, fixes the non-significant loadings from fullcfa6b2 at zero. Parameter estimates for this model are found in column 2 of Table 10 as well as in Figure 8. In this model, the highly subjective measure of student achievement attitude, sachatt, loads only on Administrative Optimism, not on Student Engagement. Six of the more concrete indicators of the Administrative Optimism latent variable are removed; only the eight most subjective remain (student achievement attitude, gang problem, student misbehavior problem, regard for property, morale of teachers as the scaling variable, teacher expectations, parent support for achievement, and parent involvement problem). This respecification provides only marginal improvements to fit indices (CFI, TLI, and RMSEA all improve by .002), but is accepted because all indicators prove significant at the .05

[102] The consistent directions of effects show that Administrative Pessimism may be a more appropriate name for the construct in these CFA models. The original name will be kept to maintain continuity with other models in this study.

level.[103] A comparison of the first two columns of Table 10 shows that this respecification causes little change in R^2 values or in the loadings of the first-order factors on the second-order factor.

Figure 8. Results of parsimonious Full-School Engagement Confirmatory Factor Analysis



A replication (*baseline25*) is performed with three modifications: (1) a new quarter is drawn, (2) listwise deletion is used in place of pairwise deletion, and (3) the jackknife is performed to more accurately estimate standard errors. Listwise deletion of all records missing any variables reduces the sample size from 1521 to 1430. Taken together, the new sample and the new method of handling missing data appear to have little effect on parameter

[103] Gangprb is not bolded and its path is italicized because it differs significantly from zero at the .05 level, but not the .01 level. All other indicators are significant at the .01 level.

estimates. Fit indices change by no more than one-thousandth from *fullcfa7b1*. Some R^2 values increase and some decrease, but none by more than .06. Some loadings increase and some decrease, but none by more than .039.

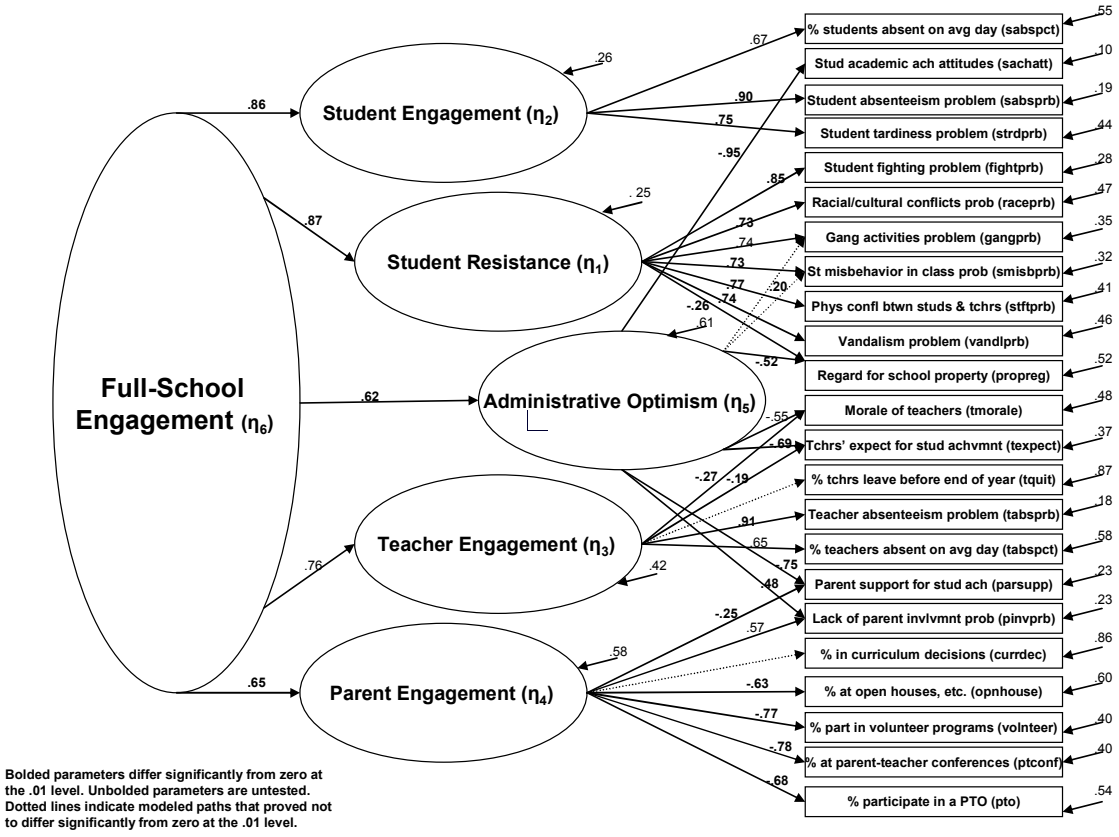
Because of NAEP's complex sampling structure, researchers are encouraged to use jackknife parameter estimates. This requires re-running the model 63 times – first with standard weights, then with the 62 different sets of replicate weights provided with the dataset. Each set of replicate weights removes some records, replacing them with duplicates of other records. The sum of the squared differences between each base model parameter estimate and its counterpart in the other 62 program runs is an estimate of the variance of the parameter. The ratio of this estimation to the estimate made under the assumptions of a simple data structure[104] is called a design effect. In this case, the design effects for the parameters are fairly small, ranging from .82 to 3.19, with a mean of 1.59. Only one parameter's significance is affected – the loading of texpect on teng becomes non-significant instead of significant at the .05 level. The jackknife variance estimates are similar to the estimates provided by Mplus.

A final replication of the best full model (baseline100) uses the full dataset, listwise deletion (N=5687), and jackknife variance estimation. This, the soundest measurement model, differs little from the previous models. It is diagrammed in Figure 9; its parameters are listed in the final column of Table 10. As with the three previous models, the fit indices (CFI, TLI, 1-RMSEA) range from .96 to .98, all strong. With the exceptions of tquit (.13) and currdec (.14), R^2 values range from .39 to .90. Prior to the jackknife, 28 of 29 loadings differ significantly from zero at the .01 level; the last (admopt by gangprb) is significant at the .05

[104] This is the estimate provided by a software program like Mplus.

level. Jackknife design effects range from 1.41 to 26.65, with a mean of 4.48. Admopt by gangprb becomes non-significant. The significance level of admopt by smisbprb and peng by currdec declines to .05. In this replication, the jackknife seems to have moderate importance, but the success of the measurement model is not altered.

Figure 9. Results of final Full-School Engagement Confirmatory Factor Analysis



Alternative models are estimated (see Table 11). The second-order factor is removed (*order1cfa3b3*). Four additional poorly-loading indicators are removed (*fullcfa9*). Parental involvement in curricular decision-making is considered as an indicator (*fullcfa10*) and predictor (*fullcfa11a*) of Full-School Engagement. All models appear to have nearly identical good fits. For the remainder of the analyses, *baseline100* is accepted as the best model on theoretical grounds.

Table 11. Alternative Confirmatory Factor Analysis models

Label	Model Name			
	order1cfa3b3	**fullcfa9**	**fullcfa10**	**fullcfa11a**
Estimator	WLSMV	WLSMV	WLSMV	WLSMV
Missing	pairwise	pairwise	pairwise	pairwise
N	1521	1521	1521	1516
Jackknife	no	no	no	no
Overall Fit Indices				
Chi-Sq	147.9	148.6	172.2	154.2
Df	45	45	45	47
P-value	0.000	0.000	0.000	0.000
CFI	0.962	0.962	0.953	0.964
TLI	0.975	0.976	0.971	0.976
1-RMSEA	0.961	0.961	0.957	0.961
Measures of Student Resistance (sres)				
~~***gangprb***~~	~~0.76~~	~~0.87~~	~~0.87~~	~~0.87~~
raceprb	**0.69**	**0.67**	**0.67**	**0.67**
smisbprb	**0.72**	**0.85**	**0.85**	**0.86**
stftprb	**0.74**	**0.72**	**0.72**	**0.73**
vandlprb	**0.78**	**0.76**	**0.76**	**0.76**
fightprb	**0.86**	**0.84**	**0.84**	**0.84**
propreg	**-0.36**	**-0.30**	**-0.30**	**-0.31**
Measures of Student Engagement (seng)				
~~***sabspct***~~	~~0.75~~	~~0.75~~	~~0.75~~	~~0.75~~
strdprb	**0.65**	**0.65**	**0.65**	**0.67**
sachatt	-0.12			
sabsprb	**0.92**	**0.91**	**0.91**	**0.92**
Measures of Teacher Engagement (teng)				
~~***tabspct***~~	~~0.68~~	~~0.68~~	~~0.68~~	~~0.69~~
tquit	**0.38**			
tabsprb	**0.90**	**0.91**	**0.91**	**0.91**
tmorale	**-0.34**	**-0.21**	**-0.21**	**-0.21**
texpect	**-0.24**			
Measures of Parent Engagement (peng)				
~~***pinvprb***~~	~~0.60~~	~~0.51~~	~~0.55~~	~~0.58~~
parsupp	**-0.26**	-0.16	**-0.16**	-0.16
opnhouse	**-0.65**	**-0.66**	**-0.64**	**-0.62**
ptconf	**-0.72**	**-0.73**	**-0.71**	**-0.70**
pto	**-0.61**	**-0.62**	**-0.60**	**-0.52**
volnteer	**-0.81**	**-0.83**	**-0.81**	**-0.79**
currdec	**-0.49**	**-0.49**		
Measures of Administrative Optimism (admopt)				
~~***tmorale***~~	~~-0.56~~	~~-0.61~~		~~-0.60~~
sachatt	**-0.87**	**-0.91**		**-0.92**
pinvprb	**0.42**	**0.48**		**0.41**

Label	Model Name			
	order1cfa3b3	fullcfa9	fullcfa10	fullcfa11a
gangprb	0.17			
raceprb				
smisbprb	**0.19**			
stftprb				
vandlprb				
fightprb				
propreg	**-0.51**	**-0.52**		**-0.51**
strdprb				
sabsprb				
tabsprb				
texpect	**-0.68**	**-0.82**		**-0.83**
parsupp	**-0.74**	**-0.78**		**-0.78**
Measures of Full-School Engagement (FSE)				
~~***teng***~~		~~0.77~~	~~0.77~~	
seng		**0.87**	**0.87**	
sres		**0.88**	**0.88**	
peng		**0.70**	**0.76**	
admopt		**0.69**	**0.68**	
currdec			**-0.38**	
Percentage of variance explained (R2)				
gangprb	0.72	0.75	0.75	0.76
raceprb	0.47	0.45	0.45	0.45
smisbprb	0.68	0.72	0.72	0.73
stftprb	0.55	0.52	0.52	0.54
vandlprb	0.61	0.58	0.58	0.58
fightprb	0.73	0.70	0.70	0.70
propreg	0.57	0.55	0.55	0.55
strdprb	0.42	0.42	0.42	0.45
sachatt	0.87	0.83	0.84	0.84
sabsprb	0.84	0.83	0.84	0.84
sabspct	0.57	0.57	0.57	0.56
tquit	0.15			
tabspct	0.46	0.47	0.47	0.48
tabsprb	0.81	0.82	0.82	0.83
tmorale	0.56	0.56	0.56	0.55
texpect	0.64	0.68	0.68	0.69
parsupp	0.78	0.75	0.76	0.76
opnhouse	0.42	0.44	0.42	0.38
ptconf	0.52	0.53	0.50	0.48
pto	0.37	0.38	0.36	0.27
volnteer	0.66	0.69	0.65	0.62
currdec	0.24	0.24	0.15	
pinvprb	0.76	0.73	0.73	0.75
sres		0.78	0.77	0.78

Label	Model Name			
	order1cfa3b3	fullcfa9	fullcfa10	fullcfa11a
seng		0.76	0.76	0.77
teng		0.60	0.60	0.62
peng		0.50	0.57	0.58
admopt		0.48	0.47	0.46
fse				0.06
Correlations				
sdis / sres	0.76			
tdis / sres	0.72			
tdis / sdis	0.67			
pdis / sres	0.60			
pdis / sdis	0.66			
pdis / tdis	0.60			
Correlations with Administrative Optimism				
sres	0.46			
sdis	0.49			
tdis	0.34			
pdis	0.44			

Note. All parameter estimates are standardized. Bold font represents parameter estimates that differ significantly from zero at the .01 level. Standard font represents estimates that differ significantly from zero at the .05 level. Italics are used for parameter estimates that do not differ significantly from zero. Strikethrough font is used to represent parameter estimates that are not significance tested; these parameters are generally used to scale the factors they measure.

These four versions of Model 1 are designed to answer the first question: *Can a single second-order latent variable called Full-School Engagement measure a constellation of factors representing administrative, parent, teacher, and student engagement in the academic mission of a school?* The analyses demonstrate that the answer is "yes."

Models 2, 3, and 4: Economic and Ethnic Composition Effects

Question 2 asks: *Do the economic or ethnic compositions of a school predict that school's mean grade 8 mathematics test scores, adjusted for the ethnicity and economic level of the individual students in that school?* This question is addressed by a sequence of three models. Model 2 is a simple regression model designed to replicate findings from many previous

studies – student ethnicity and economic level predict test scores; neither fully explains the other. Model 3 is a baseline multi-level model that determines the percent of variance in student test scores that occurs between schools and the percent of variance that occurs within schools. Model 4 puts student ethnicity and economic level in the same model as school ethnic and economic composition, allowing the separation required by Question 2 but not provided by hundreds of prior studies. The individual effects of student ethnicity and economic level on within-school test score differences are measured, as are the effects of school ethnic composition and economic composition on between-school test score differences. Because they are in the same model, each relationship is reported with the other relationships controlled, allowing an answer to the central question, which addresses pure ethnic and economic composition effects.

Model 2: Baseline Regression

The first step to answering Question 2 is a replication of basic results from the multitude of prior single-level regressions. The primary purposes are to establish the independent importance of ethnicity and economic level in the prediction of grade 8 NAEP mathematics test scores and to establish a baseline against which to compare subsequent models. A baseline single-level regression of 2003 grade 8 NAEP mathematics test scores on student ethnicity and economic level with 25% of the data and a replication with the remaining 75% of the data accomplish these purposes. The results of both models are shown in Table 12. The results of the larger replication are diagrammed in Figure 10. Important elements of the study are highlighted in this text. No tests of overall fit for these models beyond the traditional R^2 test of variance explained are possible. Traditional regression models are always just-identified, leaving zero degrees of freedom for tests of overall model fit. The two models

Table 12. Model 2. Baseline regression of grade 8 mathematics test scores on student ethnicity and economic level, with replication

Label	Model name	
	Baseline regression	Replication
Estimator	MLMV	MLMV
Missing	listwise	listwise
N	33889	103422
Jackknife	no	no
Predictors of mathematics test score		
black	**-24.2**	**-23.8**
asian	3.5	4.4
hispanic	**-15.4**	**-13.6**
amind	**-18.5**	**-12.2**
other	*-5.5*	*-1.4*
lunch status	**-20.3**	**-17.7**
title I status	**-5.8**	**-9.9**
Variance components		
Total	1254	1278
Residual	942	991
Variance	312	287
R-square	0.25	0.23

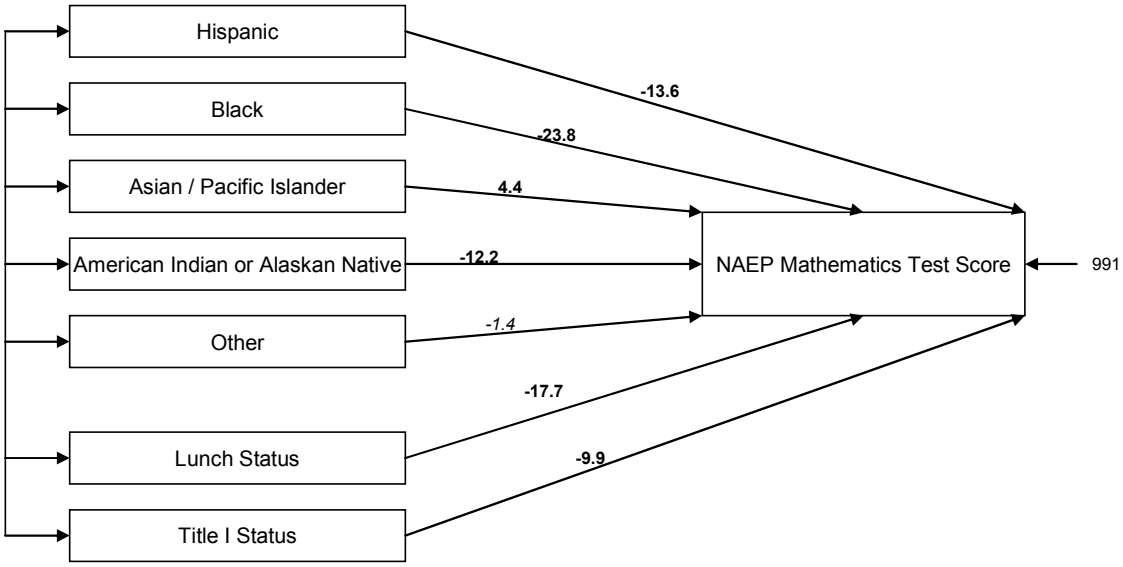
Note. The baseline regression is conducted with a random subsample of 25% of available data. The replication is conducted with the remaining 75% of data. Tests of overall model fit are not available for simple regression models. Bold font represents parameters that are significant at the .01 level. Normal font represents paramaters that are significant at the .05 level. Italics are used to represent parameters that do not differ significantly from zero at the .05 level.

predict 25% and 23% of the variance in the 2003 grade 8 NAEP mathematics scale score plausible value 1.[105] Taking the two models together and controlling for school lunch, ethnicity and Title I status, free lunch students score 18 to 20 points lower than full-price-lunch students. Reduced-price lunch students fall in the middle by design, about 9 to 10

[105] As described in Chapter 3, no single student is given the entire grade 8 NAEP mathematics assessment. For this reason, NAEP data is not reliable for the estimation of the mathematics proficiency of any given student. Researchers are provided a set of five plausible values for an individual's proficiency. The baseline regression and replication are performed with one of these plausible values, leading to an underestimation of actual test score variance and a concomittant overestimation of the significance of the reported results.

points above free lunch students and the same distance below full-price lunch students. Controlling for ethnicity and free lunch status, Title I students score 6 to 10 points lower than non-Title I students. Controlling for free lunch status and title I status, Black students score on average about 24 points lower than White students, Hispanic students about 14 points lower, American Indian students between 12 and 19 points lower, and Asian students about 4 points higher. The practical significance of these numbers is highlighted by the estimate from Chapter 3 that 11 points is equivalent to approximately one grade level between fourth and eighth grades.

Figure 10. Baseline replication of grade 8 mathematics test score regression on student ethnicity and economic level



These parameter values are a mix of individual-level and composition effects. The two levels of effects (within-school and between-school) are separated by Model 4. Nevertheless, none of these results is surprising. They simply reinforce that ethnicity and economic status predict test scores and that neither is reducible to the other. Both ethnicity and economic status dramatically affect test scores in today's U.S.

Model 3: Baseline Two-level Model

The two estimates of Model 3, a baseline two-level model, reconfirm that test scores vary both within and between schools. Firmly establishing this fact in the context of the Grade 8 NAEP dataset and estimating the percentage of total variance that occurs at each level are prerequisites for Model 4, which combines Models 2 and 3 to directly estimate the composition effects addressed by Question 2.

Table 13. Model 3. Baseline two-level model and replication. Separation of within-school and between-school variance in grade 8 mathematics test scores.

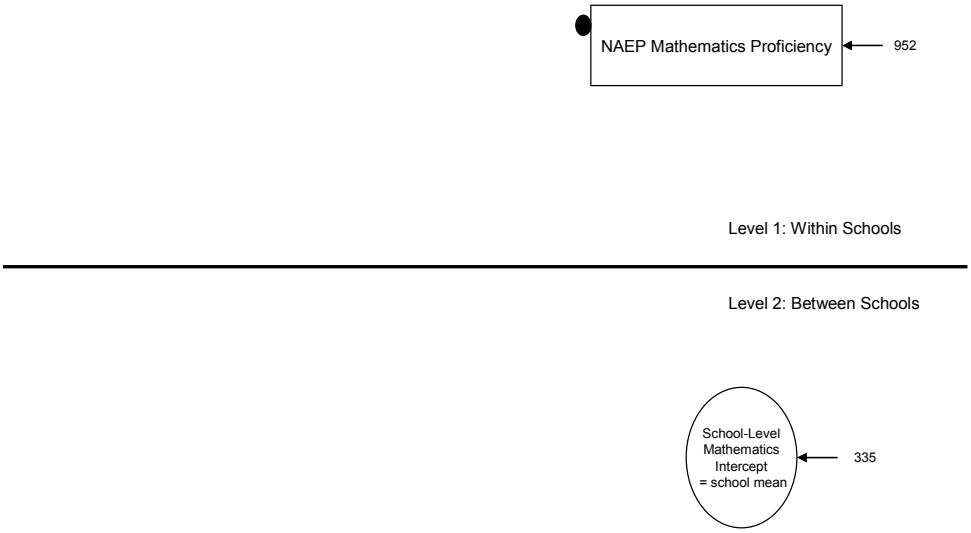
Label	Model name	
	Baseline two-level model	Replication
Estimator	MLMV	MLMV
Missing	listwise	listwise
N	38609	107681
Jackknife	no	no
Variance totals		
Within-school	898	952
Between-school	370	335
Total	1268	1287
Portion of total variation that occurs between schools		
ICC[a]	.29	.26

Note. The baseline two-level model is conducted with a random subsample of 25% of available data. The replication is conducted with the remaining 75% of data. Tests of overall model fit are not available for baseline two-level models. [a] The ICC, or intra-class coefficient provides an estimate of the portion of total variance that occurs between schools.

Model 3 is a just-identified model. With zero degrees of freedom, no tests of overall model fit are possible. The model is estimated with a random sample of 25% of the available data and replicated with the remaining 75% of the data. The results, similar across the two estimates, are shown in Table 13. Figure 11 graphically depicts the larger, and therefore

sounder, replication model. A comparison of the results with those of the baseline regression is also of interest.

Figure 11. Baseline two-level model replication. Separation of within-school variance from between-school variance in grade 8 mathematics test scores.



The total variances in the baseline regression and its replication (Model 2) are estimated at 1254 and 1278. For the baseline two-level Model 3 and its replication, these values are 1268 and 1287. The primary purpose of Model 3 is to determine what part of this variance is found within schools and what part is found between schools. The between-school variance is essentially calculated as the variance in school means, weighted by school size. The variance within each school is essentially a sum of the squared differences from the school mean. The small solid circle attached to the NAEP Mathematics Proficiency within-school variable in Figure 11 is representative of that distribution of means, the same distribution represented by the open oval in the between-schools model.

The baseline and replication models agree substantially. In the baseline model, 29% (370) of the variance is found between schools, with the remaining 71% (898) found within schools. In the replication, 26% (335) is found between schools, and the remaining 74%

(952) within. These residuals are diagrammed in Figure 11. They are considered residuals because none of the variance, in either level, is explained. Thus, Model 3 separates variance into within-school and between-school components. Between two-thirds and three-quarters of the difference in students' grade 8 NAEP mathematics test scores must be explained by individual and within-school factors; the remaining quarter to third is explained by differences between schools. Model 4 examines the degree to which individual student ethnicity and economic level explain the within-school variance and simultaneously, the degree to which school ethnic and economic composition explain the between-school variance, which might be described as overall school effectiveness.

Model 4: Composition Effects Model

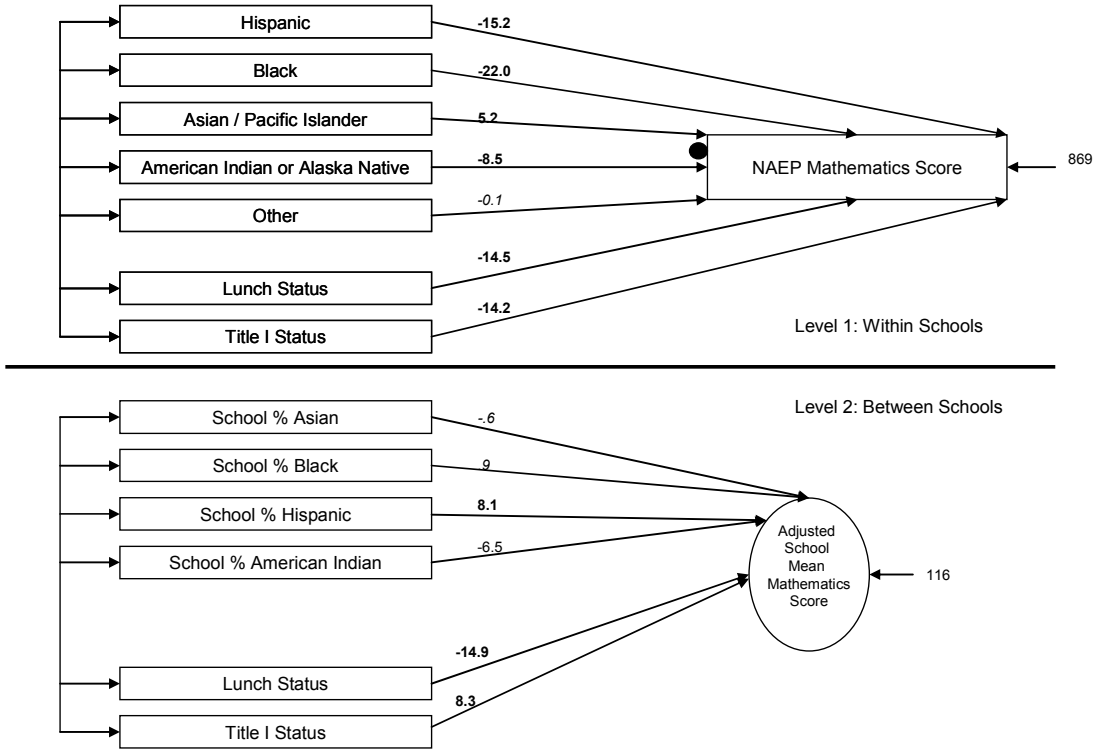
Model 2 is like almost all previous education production functions in its failure to differentiate individual and composition effects. Model 4 solves this problem by using the two-level separation of test score variance introduced in Model 3 and adding a two-level separation of the predictor variables (ethnicity and economic level) from Model 2. Student ethnicity is distinguished from school ethnic composition; student economic level is distinguished from school economic composition. The student-level variables are

Table 14. Model 4. Composition effects model. Original estimation and replication.

Label	Model name	
	Original estimation	Replication
Estimator	MLMV	MLMV
Missing	listwise	listwise
N	31248	94568
Centering	grand mean	grand mean
Jackknife	no	no
Within-school predictors of grade 8 mathematics test score		
black	**-19.5**	**-22.0**
asian	3.1	**5.2**
hispanic	**-16.8**	**-15.2**
amind	**-10.3**	**-8.5**
other	*-4.6*	*-0.1*
school lunch	**-16.5**	**-14.5**
title I	**-7.8**	**-14.2**
Between-school predictors of adjusted school mean mathematics test score		
pctblack	**-7.1**	*0.9*
pctasian	*-3.5*	*-0.6*
pcthispanic	7.23	**8.1**
pctamind	**-9.6**	-6.5
pctfl	**-20.7**	**-14.9**
pctt1	**10.4**	**8.3**
Overall variance		
Total variance	**1219**	**1278**
ICC	**.10**	**.09**
Within-school variance components		
Total Variance	**1094**	**1150**
Residual variance	**844**	**869**
Variance explained	**250**	**281**
R-square	**.23**	**.24**
Between-school variance components		
Total variance	**125**	**128**
Residual variance	**83**	**116**
Variance explained	**42**	**12**
R-square	**.34**	**.09**

Note. The baseline two-level model is estimated with a random subsample of 25% of available data. The replication is conducted with the remaining 75% of data. Tests of overall model fit are not available for this saturated model. Bold font represents parameters that are significant at the .01 level. Normal font represents paramaters that are significant at the .05 level. Italics are used to represent parameters that do not differ significantly from zero at the .05 level.

Figure 12. Model 4. Composition effects model. Replication results.



Bolded parameters differ significantly from zero at the .01 level. Unbolded parameters are untested. Dotted lines indicates modeled paths that proved not to differ significantly from zero at the .01 level.

hypothesized to predict within-school test score differences. These are within-school or individual effects. The composition variables predict between-school adjusted mean test score differences. These are between-school or composition effects. It is these composition effects that are the focus of Question 2: "Do the economic or ethnic compositions of a school predict that school's mean grade 8 mathematics test scores, adjusted for the ethnicity and economic level of the individual students in that school?"

Like Models 2 and 3, Model 4 is estimated twice – first with a 25% random sample of students, then with the remaining 75% of students. Table 14 shows results from both original estimation and replication. The results are similar, but not identical. The replication estimates, because they use the more precise larger sample, are diagrammed in Figure 12.

For ease of interpretation, the within-school variables are grand-mean centered. The advantages and details of this approach are described in Chapter 3. One result of this choice is a major shift in variance. Table 13 shows total within-school variances of 898 and 952 in the baseline two-level model and its replication. Table 14 shows corresponding values of 1094 and 1150. Table 13 shows total between-school variances of 370 and 335 in the baseline two-level model. Table 14 shows corresponding values of 125 and 128. About 200 points of variance is shifted from between schools to within schools by the act of grand-mean centering, resulting in a large drop in ICC, from .29 and .26 in Model 3 to .10 and .09 in Model 4.

These shifts are not surprising, but they do require explanation. A large share of the variance in school mean grade 8 mathematics test scores is because the "school average" test-taker in each school has a different ethnic and economic background. Grand-mean centering defines each student's ethnicity and economic level by her difference from national means. The baseline student, scoring zero on all variables, now represents a national average mix of ethnicity and economic level instead of a local average. It is in the nature of averaging that student characteristics tend to be more similar to their local school average than to the national average. Within-school variance increases because grand-mean centering tends to move most of the students in any school away from their local average.

In the same measure, between-school variance decreases because the demographic characteristics of the zero-level baseline student becomes the same for every school. It is at this zero level that the adjusted school mean is calculated. On average, the adjusted test score mean is lower than the actual mean at schools with large proportions of White, Asian, or middle-to-upper-class students and higher at schools with large proportions of Black, American Indian, Hispanic, or lower-income students.

Like Models 2 and 3, Model 4 is just-identified. There are zero degrees of freedom, therefore no tests of overall model fit are possible. The change to grand mean has some effect on within-school parameter estimates as compared to the baseline regression model (Model 2), but the overall patterns change little. Black, Hispanic, and American Indian students are still seen, on average, to have lower grade 8 mathematics test scores than the White and Asian students within their school. School lunch and Title I status continue to be significantly and negatively related to within-school scores. The within-school R^2 values remain between 23% and 25%.

The between-school results, on the other hand, hold some surprises. The percentage of Black students in a school does not have a consistently negative affect on grade 8 adjusted school mean test scores. The percentage of American Indian students does, but in the larger sample the significance level is only 5%. Counter to expectations, the percentage of Hispanic students in a school is positively related at the .05 level to grade 8 adjusted mean mathematics test scores in both original and replication models. The most consistently strong effects are those of school economic level as measured by free- and reduced-price-lunch status. With all other factors constant, a national average student in a 100% free-lunch school would be expected to score 15 to 21 points lower than the same student in a 100% full-price-

lunch school. According to model assumptions, the same composition effect applies equally to any given ethnic/economic category of students. In other words, the model assumes that school composition effects have the same impact on all groups of students.

Surprisingly, the percentage of Title I students, designed as a secondary indicator of school economic level, shows significant positive effects on adjusted school mean mathematics test scores. Controlling for all other factors (most significantly, controlling for free-lunch status), a national average student in a 100% Title I school with eighth grade students would be expected to score 8 to 10 points higher than such a student in a school with no Title I services. The between-schools R^2 is much weaker in the larger replication (.09) than in the smaller original estimate (.33), suggesting some instability in the estimation of this model.

Question 2 is answered positively with regard to economic composition effects and negatively, with some interesting exceptions, with regard to ethnic composition effects. The economic composition of a school as measured by the percentage of free-lunch students does strongly predict adjusted school mean test scores. Schools with more free- and reduced-price-lunch students tend to be less effective for all of their students. The other intended measure of the economic composition of the school turns out to have a composition effect completely opposed to that of free-/reduced-price-lunch status. Schools with higher percentages of Title I students have significantly higher adjusted mean grade 8 mathematics test scores. As will be discussed in chapter 5, this may be a measure of the effectiveness of the Title I program.

An examination of ethnic composition effects finds them to be weaker than expected. The much-discussed “Acting White” hypothesis (Fordham & Ogbu, 1986), that black student culture is damaging to achievement, is not confirmed in the larger sample. Neither does the

stereotype of the "hard-working" Asian student culture seem to have a positive effect on an entire school. Instead, the results suggest that schools with large numbers of Hispanic students are the most effective for all of their students. The closest to an expected result in the area of ethnic composition effects was for school American Indian composition. Schools with the largest shares of American Indian / Alaskan Native students were consistently the least effective when all other factors, both within and between schools, were accounted for, suggesting the possibility of a lingering anti-education culture that may be a result of the culturally genocidal origins of American Indian education in the 19th century. There are, of course, many ethnic subgroups within each of these broad ethnic groupings, including the dominant White group, which is used a baseline of comparison here. A study of these subgroups would be of interest, but NAEP provides such data only for the Hispanic group.

Model 5: Mediation Model

As described in chapter 3, the mediation model (11i4e) adds the Full-School Engagement construct to Model 4 as a mediator. This model will answer Questions 3 (*What are the relationships that exist among the economic and ethnic compositions of a school, Full-School Engagement, and adjusted school mean grade 8 mathematics test scores*?) and 4 (*Does Full-School Engagement mediate any of the composition effects identified in Question 2?*) about the role of Full-School Engagement, particularly its role as a mediator of composition effects. As with Models 2, 3, and 4, Model 5 is estimated with 25% of the data, then replicated with the other 75%. Because this is the most complete model, a third estimation is performed. This estimation uses the larger, 75%, dataset and, as described below, combines 67 model runs to make the most accurate estimates of parameters and standard errors.

Five of the runs are identical with the exception of the outcome variable. The model is run once for each of the plausible values for a student's true scale score; the parameter estimates reported for the final model are the mean of the obtained parameter estimates. The variance of these estimates is combined with jackknife variance estimates to obtain final standard errors. The results of the original estimation, its replication with the larger dataset, and the most accurate final estimation using jackknife/plausible value variance estimation are shown in Table 15. Only these final, most accurate results are diagrammed in Figure 13

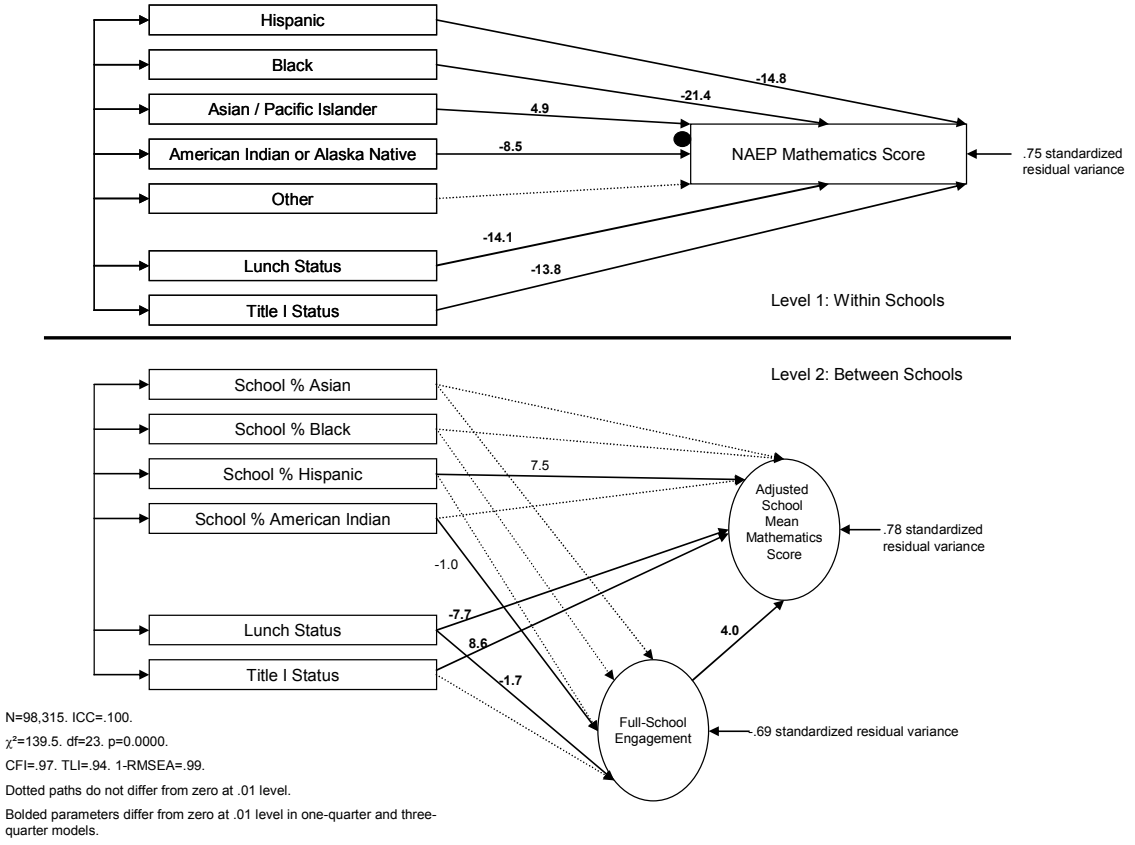
Table 15. Model 5. Mediation. Original estimation and replication.

Label	Model name		
	Original estimation	Replication	Jackknife/Plausible Value Replication
Estimator	MLR	MLR	MLR
Missing	pairwise	pairwise	pairwise
N	31248	98315	98315
Centering	grand mean	grand mean	grand mean
Jackknife	no	no	yes
Overall fit indices			
Chi-Sq	58.3	139.5	139.5
df	23	23	23
p-value	0.000	0.000	0.000
CFI	0.979	0.973	0.973
TLI	0.957	0.944	0.944
1-RMSEA	0.993	0.993	0.993
Within-school predictors of grade 8 mathematics test score			
black	**-19.5**	**-22.0**	**-21.4**
asian	*3.1*	**5.2**	**4.9**
hispanic	**-16.7**	**-15.3**	**-14.8**
amind	**-10.4**	**-8.5**	**-8.5**
other	*-4.8*	*-0.2*	*-1.2*
school lunch	**-16.5**	**-14.4**	**-14.1**
title I	**-8.0**	**-13.8**	**-13.8**
Between-school predictors of adjusted school mean mathematics test score			
pctblack	*-5.5*	*2.9*	2.1
pctasian	*-0.9*	*-0.8*	*0.5*
pcthispanic	*6.0*	**8.3**	**7.5**
pctamind	**-9.9**	*-1.9*	*-1.8*
pctfl	**-16.3**	**-8.1**	**-7.7**
pctt1	**10.6**	**8.7**	**8.6**

Label	Model name		
	Original estimation	Replication	Jackknife/Plausible Value Replication
fse	**2.4**	**4.0**	**4.0**
Between-school predictors of Full-School Engagement			
pctblack	*-0.5*	*-0.4*	*-0.4*
pctasian	*-1.1*	*-0.2*	*-0.2*
pcthispanic	*0.5*	*-0.1*	*-0.1*
pctamind	*0.3*	**-1.0**	**-1.0**
pctfl	**-1.8**	**-1.7**	**-1.7**
pctt1	*-0.1*	*-0.2*	*-0.2*
R-square for FSE	.31	.31	.31
Overall variance components			
Total variance	1220	1274	1274
ICC	.10	.10	.10
Within-school variance components			
Total	1095	1147	1147
Residual	844	869	867
Explained	251	278	280
R-square	.23	.24	.25
Between schools			
Total	125	127	127
Residual	77	98	99
Explained	48	29	28
R-square	.38	.23	.22

Note. The baseline two-level model is estimated with a random subsample of 25% of available data. The replication is conducted with the remaining 75% of data. The final model is based on 67 estimations using plausible values and jackknife variance estimation. Bold font represents parameters that are significant at the .01 level. Normal font represents paramaters that are significant at the .05 level. Italics are used to represent parameters that do not differ significantly from zero at the .05 level.

Figure 13. Model 5. Mediation. Replication results.



The complete measurement model for Full-School Engagement proves to be too computationally intensive to include in this model. Sum scores are calculated in place of maximum likelihood estimation of the four basic Engagement constructs. This simplification allows removal of Administrative Optimism from the model[106].Overall fit testing is possible in this model because the introduction of mediation and a measurement model provide some excess degrees of freedom. The fit indices are strong in all three estimations, ranging from

[106] Not surprisingly, this modified version of the Full-School Engagment construct has less explanatory power than the full version described previously. The R^2 for Student Engagement falls from .74 in the final CFA to .50 in this model, a decline of .24. R^2 declines for the other three primary engagement constructs are .20 (Teacher Engagement), .08 (Student Resistance), and .00 (Parent Engagement). The final values remain between .38 and .67.

.94 to .99. Between-school R^2 values prove somewhat more stable than with Model 4. They are .30 in the preliminary estimation and .22 in the final estimation.

With grand-mean centered within variables, the ICC is unchanged from the Composition model. However, the addition of Full-School Engagement to the model does affect the between-level estimates. No ethnic composition variable is directly predictive of grade 8 adjusted school mean mathematics test scores in both the original model and the replications. Percent Hispanic is positive in the larger models, percent American Indian negative in the smaller model. The estimates for the effect of Title I composition are almost identical to the estimates from the Composition model. The estimates for free- and reduced-price-lunch composition effects are decreased by the addition of the Full-School Engagement mediator – from 21 to 16 in the smaller sample, and from 15 to 8 in the larger sample. In both samples, Full-School Engagement significantly predicts grade 8 adjusted school mean mathematics test scores. The only consistent predictor of Full-School Engagement is the percentage of free- and reduced-price-lunch students in the school, yet 31% of the variance of Full-School Engagement is explained.

In the final model, the jackknife variance estimation has virtually no effect. The design effects range from 0.58 to 1.48, with a mean of 0.94. No parameter's significance level is affected. Final variance and parameter estimates, however, must take plausible values into account. The variance added by these values again has no effect on the significance of any variable.

Question 3 asks: *What are the relationships that exist among the economic and ethnic compositions of a school, Full-School Engagement, and adjusted school mean grade 8 mathematics test scores?* Models 4 and 5 suggest that schools with more low-income

students tend to be less effective for all their students and also to have lower Full-School Engagement; that schools with more Title I students tend to be more effective than they otherwise would be; and that Full-School Engagement is an important predictor of school effectiveness. The final question, *Does Full-School Engagement mediate any of the composition effects identified in Question 2?* is discussed in the next section.

Full-School Engagement as a Partial Mediator of Composition Effects

In the largest, most fully specified model (Model 5), Full-School Engagement is confirmed as a partial mediator of the effect of economic composition on grade 8 adjusted school mean mathematics test scores. The direct composition effect of moving a student from a 100% free-lunch school to a 0% free-lunch school is 14.9 NAEP scale score points in the model without the Full-School Engagement mediator (Model 4), but only 7.7 points in the model with the Full-School Engagement mediator (Model 5).

This decline in the direct effect is because of the introduction of an indirect effect. In the most complete estimation, a move from a 100% free-lunch school to a 0% free-lunch school leads, on average, to a 1.7 standard deviation increase in Full-School Engagement. A 1.0 standard deviation increase in Full-School Engagement causes, on average, a 4.0 point increase in adjusted school mean grade 8 mathematics test score. Multiplication of these two parameters yields a value of 6.8, the indirect effect of school economic composition on school adjusted mean test scores via Full-School Engagement. The sum of the direct and the indirect effects (7.7 + 6.8) is 14.5, not far from the direct estimate in Model 4. Full-School Engagement is a classic partial mediator of the effect of school economic composition on adjusted school mean test scores. In other words, part of the reason that students in schools

with more low-income students have lower test scores is that such schools are characterized by lower levels of engagement by all parties in the school.

These models suggest that Full-School Engagement may also be a mediator of an ethnic composition effect. Model 4 suggests that schools with larger percentages of American Indians/Alaskan Natives tend to have lower test scores for all of their students, even controlling for the ethnicity and economic levels of those students. Table 14 shows estimates of 9.6 and 6.5 points as the effect of a shift from a 100% American Indian school to a 0% American Indian school. Despite relatively small sample sizes for this population, estimates are significant at the .05 level. Table 15, which includes the Full-School Engagement mediator, provides quite different estimates of the direct American Indian composition effect in the different samples. The coefficient is a significant 9.9 with the smaller preliminary sample, and an non-significant 1.8 in the final estimation. Accepting the final estimate raises the possibility that the entire effect of American Indian composition is indirect – by way of Full-School Engagement. A significant coefficient of -1.0 for the path from pctamind to Full-School Engagement combines with a significant coefficient of 4.0 for the path leading from Full-School Engagement to adjusted school mean test score, suggesting an indirect effect of about 4.0 points. We might conclude that lower levels of engagement by all parties to the schooling process in schools with large percentages of American Indians / Alaskan Natives lead to lower test scores for all students in the school. This inference is weakened, however, by the original estimation with the 25% sample, in which the path from pctamind to fse is not significant[107] and the path from pctamind to test score is significant and large. The instability

[107] With the 25% sample and the relatively small ethnic group, insignificance may be due to low statistical power.

of these results indicates some support for the American Indian/Full-School Engagement mediation hypothesis, but the results are clear for school economic composition. It is fair to say that these models provide support for the hypothesis that schools with more low-income students tend to be characterized by lower levels of engagement by all parties to the schooling process, and that this is part of the reason for lower mean scores in those schools. More discussion follows in the summary and in chapter 5.

Summary

This study produces three major expected results. First, it is possible to successfully operationalize a Full-School Engagement construct using NAEP data (Question 1). Second, 2003 grade 8 NAEP mathematics test score variance occurs both between and within schools. Third, Full-School Engagement partially mediates the composition effect of school economic level on grade 8 adjusted school mean mathematics test scores.

Unexpectedly, no consistent ethnic composition effects are identified, although a negative effect for school percentage American Indian (mediated by Full-School Engagement) and a positive effect for school percentage Hispanic (not so mediated) are suggested. Also unexpectedly, a consistent positive composition effect is found for the percentage of Title I students, suggesting a possible validation of the effectiveness of this program, but more investigation is needed. Methodologically, jackknife variance estimation appears to add only a little to two-level variance estimation, such as is available in the Mplus program. The results are not strong enough to recommend against the jackknife, but further investigation of the necessity of this laborious variance estimation method may be warranted.

CHAPTER FIVE - DISCUSSION

The purpose of this study was to investigate the degree to which Full-School Engagement explains the grade 8 mathematics test score gaps that exist between economically and ethnically differing groups of students. The investigation addressed the following four questions:

Question 1. Can a single second-order latent variable called Full-School Engagement measure a constellation of factors representing administrative, parent, teacher, and student engagement in the academic mission of a school?

Question 2. Do the economic or ethnic compositions of a school predict that school's mean grade 8 mathematics tests scores, adjusted for the ethnicity and economic level of the individual students in that school (i.e., composition effects)?

Question 3. What are the relationships that exist among the economic and ethnic compositions of a school, Full-School Engagement, and adjusted school mean grade 8 mathematics test scores?

Question 4. Does Full-School Engagement mediate any of the composition effects identified in Question 2?

To address these questions, a sequence of five research- and theory-based models was proposed. Each of the models was estimated with a preliminary and then a more complete sample from the full 2003 NAEP grade 8 mathematics database. Model 1 was a Confirmatory

Factor Analysis of the Full-School Engagement construct, designed to address Question 1. The construct was successfully defined. Full-School Engagement was found to be reasonably viewed as a second-order factor describing a school and influencing the levels of Student Engagement, Student Resistance, Teacher Engagement, Parent Engagement, and Administrative Optimism. Administrative Optimism served both as a methodological factor designed to partially explain shared variance among the many subjective items in the administrative survey and as one of the measures of Full-School Engagement. Only the most subjective items loaded strongly on Administrative Optimism, prompting the one respecification in this study.

Model 2 (Question 2) was, in the tradition of education production functions, a simple regression of test scores on student ethnicity and economic level. The results were as expected:

- Controlling for student ethnicity, free/reduced-price lunch students and Title I students had lower mathematics scores than their peers.
- Controlling for student economic level, Black, Hispanic, and American Indian students scored more poorly than their White and Asian peers.

Model 2 suffered from the same flaw as the large majority of its production function predecessors: it failed to distinguish within-school and between-school factors. Models 3 and 4 were designed to solve this problem.

Model 3 (Question 2) was a null two-level model. With no predictors, it simply divided grade 8 NAEP mathematics test score variance into two parts. Approximately one quarter of the variance was found between school means and approximately three-quarters of the variance was found around the means within schools.

Model 4 (Questions 2 and 4) was designed to estimate economic and ethnic composition effects on NAEP grade 8 mathematics test scores. The test score outcome variable was treated at two levels as with Model 3. The predictors from Model 2 were also separated into two levels, and all variables were grand-mean centered. The school mean test scores used to calculate the between-school variance in Model 2 were replaced with school means adjusted to represent the predicted score for a student with national-average (rather than school-average) demographics. These adjusted school means vary much less than the actual school means, creating a dramatic reduction in between-school variance. This variance does not disappear; it becomes within-school variance because each student's score is now compared with the score of a student with national-average, rather than school-average, demographics. This adjusted average student may be far from the center of the actual school distribution in terms of demographics and test scores, increasing within-school variance. A result, then, of this grand-mean centering choice was that within-school variance increased to about 90% of overall variance, with only about 10% of variance between schools. Another way to explain this large shift in variance is that 75% of test score variance is seen within schools, 15% is seen between schools because different schools have different kinds of students, and the remaining 10% is because of the differential effectiveness of different schools. Grand-mean centering, then, is a way to separate out the variance that speaks to school effectiveness. The focus of Model 4 was to begin the explanation of that variance.

The focus of Model 4 was to what degree the ethnic and economic characteristics of schools could explain the 10% of variance that represented differences in school effectiveness. In addition, the model estimated the effects of individual ethnicity and economic level on the other 90% of variance, considered to be within schools in the grand-

mean centered model. The within-school results matched the results from the single-level regression model fairly closely. Both ethnicity and economic level were important predictors of test scores in well-known ways, each controlling for the other.

The between-school results were more interesting because they moved into less well-known territory. They spoke to the question of composition effects. Do the ethnic and economic compositions of schools relate to the effectiveness of those schools? The model made a simplifying assumption that any such effects would be the same for all groups of students in the school. A strong result was found in both preliminary and final models for the composition effect of school economic composition as measured by the percentage of free-lunch students in the school. Under this model and controlling for all other school characteristics, a student attending a 100% free-lunch school had, on average, a 15 to 21 point worse NAEP scale score than a similar student attending a 0% free-lunch school. This is the equivalent of one to two grade levels, a profound effect.

The percentage of Title I students in the school was also a strong predictor of test scores, but in the opposite direction from that expected. Controlling for student ethnicity and economic level and for school ethnic composition and free-lunch composition, schools with more students in Title I programs tended to have significantly higher test score averages than schools with less students in Title I programs. This composition effect is strong and in the opposite direction from the individual-level effect. One interpretation would be that Title I students have weaker skills than their non-Title I peers, but schools with a lot of the support that comes with Title I dollars are more effective than schools with less Title I support. This finding is worthy of further investigation.

There were no consistent ethnic composition effects at the .01 level. Schools with more Black students were less effective for all students in the preliminary model, but the sign was reversed and the result was non-significant in the final model. The percentage of American Indians or Alaskan Natives in the school was related to lower effectiveness for all groups of students in the school in both models. This result was significant at the .01 level in the preliminary model and at the .05 level in the final model. The composition effect for Hispanic students was significant at the .05 level in the preliminary model and at the .01 level in the final model. Like Title I status, the sign for the composition effect was reversed from the sign for the individual effect. While Hispanic students are, on the whole more challenged mathematically than their non-Hispanic peers, schools with large Hispanic populations apparently have one or more attributes that partially overcome the challenges these students face. Perhaps schools with a critical mass of Hispanic students make the changes needed to serve them effectively, such as finding sufficient numbers of bilingual teachers. Or perhaps the Hispanic community positively affects schools in which it is strongly represented. The percentages of Asian students in a school had no relationship with school effectiveness in either model.

Model 5 (Questions 3 and 4) retained all of the paths from Model 4, while adding Full-School Engagement as a mediator of the composition effects described above. Because of computational power constraints, simultaneous estimation of the full second-order factor model and the associated path model was not possible. Sum scores were used to represent Student Engagement, Student Resistance, Teacher Engagement, and Parent Engagement. These constructs were then used as measures of Full-School Engagement. The central result of the study supported the hypothesis that Full-School Engagement acts as a mediator of

economic composition effects. Schools with larger numbers of low-income students tend to have lower test scores, partly because all the parties involved in the schooling process tend to be less engaged than they would be at schools with more well-to-do students. In a more tenuous conclusion, schools with more American Indian students may suffer similar weakness in Full-School Engagement, leading to lower test scores.

Conclusions

The first question of this study was: "Can a single second-order latent variable called Full-School Engagement measure a constellation of factors representing administrative, parent, teacher, and student engagement in the academic mission of a school?" Model 1 demonstrated that the answer to this question is yes. It is reasonable to think of Full-School Engagement as a property of a school that encompasses the engagement of parents, teachers, and students as well as student resistance and administrative optimism. This supports some research in the field of school climate, but a new term is used because of the wide variation in the operationalizations of school climate in prior research.

The second question of the study was: *Do the economic or ethnic compositions of a school predict that school's mean grade 8 mathematics tests scores, adjusted for the ethnicity and economic level of the individual students in that school (i.e., composition effects)?* Taken together, Models 2, 3, and 4 suggest complex answers to these questions. Controlling for all other individual and school-level demographic variables,

- Schools with higher proportions of free- or reduced-price-lunch students appear to be less effective.
- Schools with higher proportions of Title I students appear to be more effective.

- Schools with higher proportions of Black or American Indian students may be less effective.
- Schools with higher proportions of Hispanic students may be more effective.

Clearly, there is still much to be learned about the complexity of ethnic composition effects.

The third question of the study was *What are the relationships that exist among the economic and ethnic compositions of a school, Full-School Engagement, and adjusted school mean grade 8 mathematics test scores?* Model 5 suggests that raising levels of Full-School Engagement might significantly improve adjusted school mean mathematics test scores and that schools with more free- and reduced-price-lunch students tend to have lower Full-School Engagement. The final model suggests that schools with more American Indian students also tend to have lower Full-School Engagement, but this conclusion of the final replication model was not supported by the preliminary model.

The fourth question of the study was *Does Full-School Engagement mediate any of the composition effects identified in Question 2*? Models 4 and 5, taken together, suggest that Full-School Engagement partially mediates the composition effect of school economic level on NAEP grade 8 mathematics test scores. The addition of Full-School Engagement to the model added significant paths from Full-School Engagement to test scores and from school economic composition to Full-School Engagement. Furthermore, it reduced the direct effect of school economic composition on test scores. This set of facts implies that **schools with more low-income students tend to have lower Full-School Engagement and that this partly explains why they are less effective at improving the overall test scores of their students**. This is the key conclusion of the study.

Recommendations for Educational Policy

What are the implications for educational policy that economic, but not ethnic, composition effects appear to partly explain test score gaps and that differential Full-School Engagement partly explains the economic composition effects? The findings are relevant to discussions about how to close test score gaps. They support efforts to desegregate schools along economic lines and, failing that, efforts to improve multiple-dimensional communities in low-performing, low-income schools.

Desegregation efforts in the U.S. have historically focused on race. The legally enforced segregation of Black and White students in Southern schools though the 1950s is now universally seen as a moral outrage. Yet de facto racial segregation still exists and is, in fact, increasing (Boger & Orfield, 2005; Orfield, 2001; Orfield & DeBray, 1999; Orfield & Yun, 1999; Rumberger & Palardy, 2005). This is a problem for our nation for many reasons, but this study suggests that if our only goal is to improve test scores, it may be more important to desegregate along economic than along ethnic lines.

Because of the close links between ethnicity and economic level in the U.S., decreasing one form of segregation almost always decreases the other, but the choice of focus can be important. Wake County, NC, has shown that a magnet-school based plan of economic desegregation can be effective at maintaining relative equity within a school system (Regan, 2005), but within-district desegregation has limited effectiveness because between-district segregation is a far more powerful force than within-district segregation and much harder to overcome (Orfield, 1996). Nevertheless, this study provides evidence that overcoming economic segregation is very important, supporting the strong evidence provided by Willms

(2006) that socioeconomic desegregation is a key tool for achieving more equitable outcomes in every nation of the world.

Desegregation is a very important goal for policymakers to strive for, but this study also provides support for one approach to improving test scores at an individual school level. The relationship between school economic level and test scores is partly because high poverty schools tend to have low levels of Full-School Engagement. The success of Comer's School Development Project may be due, in part, to successful efforts to break that link (G. D. Borman et al., 2003; Comer, 2004; Comer et al., 1996; Noblit, Malloy, & Malloy, 2001). A community that includes parents, teachers, administrators, and, to a lesser degree, students is intentionally developed in these schools, the majority of which serve low-income students. The test score results in Comer schools approach those of Comprehensive School Reform models, which focus much more directly on curriculum and tests (G. D. Borman et al., 2003), but the benefits may range much more widely.

As important as the theories and approaches that are supported by this study, are those that are not supported. Cultural dissonance and "Acting White" hypotheses would suggest a powerful ethnic composition effect, particularly for Black students. This effect was not found, suggesting that these theories may not be valid, that positive attitudes toward the education system in the Black community may balance the negative effects of cultural dissonance, or that segregation might have the effect of building positive Black community (hooks, 1994; V. E. Lee et al., 1993).

Recommendations for Future Research

This study is just the first piece of a NAEP-based research agenda that attempts to uncover the patterns in the NAEP data that explain the reasons for the test score gaps that plague our nation's educational system. Future studies can

- look more carefully at the various family characteristics that are often lumped together as socioeconomic status (Bollen et al., 2001) and how each of them relates to educational outcomes and mediators of educational outcomes, with the goal of understanding theoretically and empirically how family characteristics are related to each other and to test score outcomes;
- look more carefully at subgroupings of Hispanic students;
- consider geographic variables such as region and community type (rural, urban, suburban);
- consider the role of school sector – are various kinds of private and charter schools more effective than (Gamoran, 1992; K. A. Johnson, 2000; V. E. Lee & Bryk, 1988, 1989), less effective than (C. Lubienski & Lubienski, 2006), or about the equal of (Braun, Jenkins, & Grigg, 2006) public schools? Do these differences have an effect on ethnic and economic test score gaps?
- test Bourdieu's theories regarding cultural capital (Bourdieu, 1973, 1986; Driessen, 2001; J.-S. Lee & Bowen, 2006; Sullivan, 2001);
- investigate further the possibility of using NAEP data to verify the effectiveness of programs such as Title I (G. D. Borman, 2005; G. D. Borman & D'Agostino, 1996; Orfield & DeBray, 1999; Schellenberg, 1999) and special education (Artiles, Klingner, & Tate, 2006; Harry & Klingner, 2006; Y. Perry et al., 2000);

- investigate the roles of tracking (Darling-Hammond, 2006; Loveless, 1999; Nind, Rix, Sheehy, & Simmons, 2005; Sleeter, 2005; Spade, Columbia, & Vanfossen, 1997), teacher quality (Darling-Hammond, 2000, 2002-2003, 2005; Education Trust - West, 2005; King, 2006), and reform orientation (S. T. Lubienski, 2006; Mayer, 1999; Wenglinsky, 2002, 2004) in the generation of educational inequality;
- bring all these pieces together to examine the reproduction hypothesis (Bourdieu & Passeron, 1990): put colloquially, "them as got, gets."

Methodologically, future studies can be improved by

- the ability to estimate more complex models using higher-powered computers (e.g. simultaneous estimation of the FSE factor model and the full, final path model);
- investigating alternative models, such as the possibility that Full-School Engagement is a result of, rather than a cause of, high test scores and the possibility that the various forms of Engagement have such different relationships with test scores, ethnicity, and economic level that they are better modeled as correlated constructs than as measures of a second-order construct;
- modeling of interaction effects such as the possibility that schools may be differentially effective for different groups of students;
- considering non-linear relationships among variables;
- repeating analyses on large datasets representing different populations (Willms, 2006);
- repeating analyses with different outcome variables, such as reading scores and graduation outcomes;

- repeating analyses on datasets with pre-tests to overcome the difficulty with deriving causal conclusions from cross-sectional data.

Final Remarks

This study has brought together a number of powerful tools – NAEP's restricted use database, structural equation modeling**,** multi-level modeling, and simultaneous modeling of ethnic and economic variables – to test and estimate a well-theorized model of one of the reasons for the large differences in test scores between eighth grade students from different economic levels and ethnic groups. Some of the hypotheses were supported; some were not. It is hoped that a greater understanding of our nation's educational system is one result and that the value of the methods used has been demonstrated.

The rich data collected by the NAEP program, with its strong measure of mathematics ability, its large collection of background variables, its psychometric quality, and its growing importance as a proto-national assessment, deserves many more methodologically sophisticated studies such as this one. Structural Equation Modeling's tools for improving measurement of constructs like Full-School Engagement and for modeling mediation are shown to be applicable even to a dataset as complex as NAEP.

The importance of multilevel modeling of test score data and of the relationships of predictors with test scores is reconfirmed. But it is hoped that readers will gain a deeper appreciation for the ways that multi-level analysis can separate within-school variation from between-school variation, the ways that centering of variables can dramatically alter the meanings of the two kinds of variation and separate the variation due to school effectiveness from variation due to the characteristics of the average student in a school, and the ways that relationships among variables may differ greatly within and between schools.

Inclusion of ethnic and economic variables at both within-school and between-school levels of the model proved to be important. Neither variable is sufficient to explain test score variance by itself. Between-school relationships are not necessarily the same as within-school relationships. For example, even controlling for economic level, Hispanic students tend to score more poorly than White and Asian students on grade 8 NAEP mathematics assessments, yet schools with large percentages of Hispanic students are more effective than those with less Hispanic students.

All of these matters may seem technical, but it is hoped that this study shows the practical importance of such distinctions. Future studies of the effects of various policies, teaching methods, and resource allocations should incorporate the methods demonstrated here to the greatest degree possible.

APPENDIX A - EQUATIONS AND MATRICES

This study consists of a sequence of five structural equation models. The equations and matrices associated with each model are presented in this appendix. Structural equation models can be described with a pair of interrelated submodels – a measurement model and a structural model (Heck, 2001). The general equation of the measurement model used for this study is

$$y_i = \nu + \Lambda\eta_i + \varepsilon_i$$

where y_i is a vector of dependent variables observed for individual i, ν is a vector of measurement intercepts, Λ is a matrix of measurement slopes, η_i is a set of latent variables, and ε_i is a vector of residuals uncorrelated with other variables. The covariance matrix of ε_i is denoted θ. The general equation of the structural model used for this study is

$$\eta_i = \alpha + B\eta_i + \Gamma x_i + \zeta_i$$

where η_i is the same set of latent variables contained in the measurement model, α is a vector of latent variable intercepts, B is a matrix of regression coefficients relating factors to one another, Γ is a matrix of regression coefficients relating the exogenous x_i variables to the latent variables, and ζ_i is a vector of residuals indicating that the endogenous factors are not perfectly predicted by the structural equations. The covariance matrix of ζ is denoted ψ. For the remainder of this appendix, the individual (i) subscript will be dropped from variable vectors.

Model 1. Full-School Engagement CFA.

The Full-School Engagement confirmatory factor analysis (CFA) includes five first-order latent variables (snonres, seng, teng, peng, admopt) and one second-order latent variable (fse).

$$\eta^T = [\ \text{snonres seng teng peng admopt fse}\ ]$$

The first-order latent variables are measured by a collection of 23 ordinal y-variables. Tquit is a seven-category variable; the others all have four categories.

$$y^T = [\ \text{gangprb raceprb smisbprb stftprb vandlprb fightprb propreg strdprb sachatt sabsprb sabspct tquit tabspct tabsprb tmorale texpect parsupp opnhouse ptconf pto volunteer currdec pinvprb}\ ]$$

A summary of categorical data proportions shows all of the *y*-variables to be skew. As a part of the estimation process, 72 threshold values ($\tau_{1\text{-}72}$: 6 for tquit and 3 for each of the other 22 variables) are estimated. These threshold values are the cutpoints on presumed underlying continuous variables that define the values of the categorical variables. The relationships of these continuous underlying variables with other variables in the model can be analyzed using standard statistical tools and formulas. The ν matrix provides

intercepts for these underlying variables. The Mplus default parameterization sets the intercepts of latent variables such as these underlying normal variables to 0[108].

ν^T = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

Λ is a 23 by 6 matrix of measurement slopes. Each row represents an observed *y*-variable. Each column represents a latent variable. In a confirmatory factor analysis such as this one, the majority of the loadings are set to zero, representing the investigator's hypothesis that most observed variables are unrelated to most latent variables. The λ values represent the loadings (or regression slopes) of the *y*-variables on the latent variables. Each latent variable needs a scale. This is commonly done by fixing the latent variable to the value of one of its measures. The 1 values in the Λ matrix indicate that admopt is fixed to the value of morale of teachers and snonres is fixed to the value of gangprb. The -1 values in the Λ matrix indicate that seng, teng, and peng are fixed to the opposites of the values of sabspct, tabspct, and pinvprb, respectively. The first four columns represent a traditional simple confirmatory factor analysis, with no measures shared by multiple latent variables. The fifth column represents admopt, a variable that shares measures with each of the first four variables.

---

[108] An equivalent parameterization would set the first two thresholds to zero and one, allowing estimation of the mean and variance of the underlying variables, but these estimations are not central to this study.

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 & 0 & \lambda_{1,5} & 0 \\ \lambda_{21} & 0 & 0 & 0 & \lambda_{2,5} & 0 \\ \lambda_{31} & 0 & 0 & 0 & \lambda_{3,5} & 0 \\ \lambda_{41} & 0 & 0 & 0 & \lambda_{4,5} & 0 \\ \lambda_{51} & 0 & 0 & 0 & \lambda_{5,5} & 0 \\ \lambda_{61} & 0 & 0 & 0 & \lambda_{6,5} & 0 \\ \lambda_{71} & 0 & 0 & 0 & \lambda_{7,5} & 0 \\ 0 & \lambda_{82} & 0 & 0 & \lambda_{8,5} & 0 \\ 0 & \lambda_{92} & 0 & 0 & \lambda_{9,5} & 0 \\ 0 & \lambda_{10,2} & 0 & 0 & \lambda_{10,5} & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_{12,3} & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_{14,3} & 0 & \lambda_{14,5} & 0 \\ 0 & 0 & \lambda_{15,3} & 0 & 1 & 0 \\ 0 & 0 & \lambda_{16,3} & 0 & \lambda_{16,5} & 0 \\ 0 & 0 & 0 & \lambda_{17,4} & \lambda_{17,5} & 0 \\ 0 & 0 & 0 & \lambda_{18,4} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{19,4} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{20,4} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{21,4} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{22,4} & 0 & 0 \\ 0 & 0 & 0 & -1 & \lambda_{23,5} & 0 \end{bmatrix}$$

All of these variables are based on responses by a single administrator to a single survey. Shared variance, particularly on the more subjective items, may indicate the optimism of that administrator as much as the engagement of the parties involved in the schooling effort. This administrative optimism is also seen as one of the indicators of full-school engagement. The sixth column represents the second-order latent variable, full-school engagement. It is hypothesized to be directly related to none of the observed

variables. Its indirect relationship with them is contained in the structural model, not the measurement model.

The α vector contains the intercepts of the latent variables. In this case, all are fixed to zero.

$$\alpha^T = [\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

B is a matrix of regression coefficients relating the six factors to each other. The only relationship proposed in this model is that the second-order factor, Full-School Engagement, predicts the other five factors. The final column of the matrix contains the regression coefficients (also known as loadings) that represent these relationships. The 1 in row three shows that fse is set on the same scale as teng. As shown in the matrix, all other relationships between these latent variables are presumed to be zero.

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \beta_{16} \\ 0 & 0 & 0 & 0 & 0 & \beta_{26} \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & \beta_{46} \\ 0 & 0 & 0 & 0 & 0 & \beta_{56} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Because there are no exogenous x-variables in this model, there is no Γ matrix. Each school in this analysis has values for snonres, seng, tend, peng, and admopt that are imperfectly predicted by fse. ζ_i is a six by one vector containing these five residuals and the value of fse for school i. ψ is the six by six covariance matrix of these ζ_i. It is a diagonal matrix because the ζ_i are presumed independent.

$$\text{Diag } \psi = [\ \psi_{11}\ \psi_{22}\ \psi_{33}\ \psi_{44}\ \psi_{55}\ \psi_{66}\ ]$$

The Θ matrix contains the variances and covariances of $\varepsilon_{1\text{-}23}$. It is a 23 x 23 matrix, diagonal because these error variances are presumed independent of each other.

$$\text{Diag } \Theta = [\ \theta_{11}\ \theta_{22}\ \theta_{33}\ \theta_{44}\ \theta_{55}\ \theta_{66}\ \theta_{77}\ \theta_{88}\ \theta_{99}\ \theta_{10,10}\ \theta_{11,11}\ \theta_{12,12}\ \theta_{13,13}\ \theta_{14,14}\ \theta_{15,15}\ \theta_{16,16}\ \theta_{17,17}\ \theta_{18,18}\ \theta_{19,19}$$

$$\theta_{20,20}\ \theta_{21,21}\ \theta_{22,22}\ \theta_{23,23}\ ]$$

The data source for a structural equation model is the covariance matrix of the observed variables – or, in this case, the presumed underlying normal versions of the observed variables. This model includes 23 variables. The covariance matrix therefore includes 23 x 24/2 = 276 data points. The model estimates 72 threshold values, 33 first-order loadings, 4 second-order loadings, 23 observed variable error variances, 5 latent variable residual variances, and the variance of the single second-order latent variable, fse. This is 138 estimated parameters.

Weighted least square parameter estimates are obtained using a diagonal weight matrix with standard errors and mean- and variance-adjusted chi-square test statistics that use a full weight matrix. Mplus software is used to perform these calculations with the WLSMV (e.g. DWLS) estimator. A conventional maximum likelihood analysis would have 276 – 138 = 138 degrees of freedom for this analysis, but a distinct Mplus formula estimates the degrees of freedom more accurately to be 44 for this model.

Non-negative degrees of freedom satisfy a necessary condition (the t –rule) for the identification[109] of this model. None of the sufficient conditions for estimation are met because there are two non-zero elements on some rows of Λ (Bollen, 1989, p. 247). For this reason, the identification of the model is checked empirically.

Model 2. Baseline Regression.

The baseline regression model was estimated with standard OLS techniques and, equivalently, as a structural equation model with no latent variables. Viewed as an SEM, it uses the same pair of matrix equations listed in the first paragraph of this technical appendix:

$$y_i = \nu + \Lambda\eta_i + \varepsilon_i$$

$$\eta_i = \alpha + B\eta_i + \Gamma x_i + \zeta_i$$

In this formulation, all of the variables in the model are viewed as y-variables.

y = [mrpcm1 title1w black asian hispanic amind other rlunchw]

In a standard OLS formulation, only mrpcm1, the NAEP mathematics proficiency variable, would be considered a y-variable. The intercepts of the variables are all fixed at zero.

$$\nu^T = [\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

Each variable is treated as a single, perfect indicator of a corresponding latent variable.

$$\eta^T = [\ \text{mrpcm1 title1w black asian hispanic amind other rlunchw}\ ]$$

[109] Identification of a structural equation model is equivalent to the theoretical estimability of that model (Bollen, 1989).

$$\Lambda = \text{diag}\,[\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ ]$$

$$\Theta = \mathbf{0}$$

The intercept of each latent variable is fixed at the mean of the corresponding observed variable in the sample which is the mean of the weighted observed variable. Only the intercept of mrpcm1 is estimated.

$$\alpha^T = [\ \alpha_1\ \ .268\ .162\ .032\ .114\ .017\ .005\ .342\ ]$$

All of the relationships between the latent variables are fixed at zero except the seven estimates of the regression of mrpcm1 on title1w, black, asian, hispanic, amind, other, and rlunchw.

$$B = \begin{bmatrix} 0 & \beta_{12} & \beta_{13} & \beta_{14} & \beta_{15} & \beta_{16} & \beta_{17} & \beta_{18} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

All of the variances and covariances among the latent variables are fixed except the residual variance of the outcome variable – mrpcm1. The variances and covariances of the seven independent variables are fixed at weighted sample values. The covariance of each of these IVs with the DV (mrpcm1) is set to zero because it is estimated as a regression coefficient in the B matrix.

$$\Psi = \begin{bmatrix} \psi_{11} & & & & & & & \\ 0 & .196 & & & & & & \\ 0 & .040 & .136 & & & & & \\ 0 & .001 & -.005 & .031 & & & & \\ 0 & .033 & -.018 & -.004 & .101 & & & \\ 0 & .004 & -.003 & -.001 & -.002 & .017 & & \\ 0 & .000 & -.001 & .000 & -.001 & .000 & .005 & \\ 0 & .079 & .049 & .001 & .037 & .004 & .000 & .204 \end{bmatrix}$$

Regression models like this one are just-identified, with zero degrees of freedom, and always identified. This model is estimated using the Mplus MLMV estimator. The parameters are estimated with maximum likelihood. Standard errors are robust to non-normality. No chi-square test statistics are available for saturated models.

Model 3. Baseline two-level.

The goal of the two-level analysis in this study is decompose the variance in mathematics test scores into within-school and between-school components and then to use a set of predictors to explain the variance present at each level simultaneously. For each student, the total score is decomposed into a between component (the school mean or adjusted mean) and a within component (the deviation of the student's score from the school mean or adjusted school mean). This individual decomposition allows the maximum likelihood estimation of separate within- and between-groups covariance matrices. These matrices then provide the basis for providing optimal parameter estimates at both levels simultaneously using the same equations provided at the beginning of this appendix (Heck, 2001, p. 101).

In the baseline two-level model, there are no predictor variables. Only three parameters are estimated: the variance of mrpcm1 within schools (θ_{w11}), the variance of mrpcm1 between schools (θ_{b11}), and the mean of mrpcm1 between schools (ν_{b1}). The within-school mean (ν_{w1})is fixed at zero.

The intra-class correlation (ICC) can be calculated as $\rho = \frac{\theta_{b11}}{\theta_{b11} + \theta_{w11}}$. The ICC indicates the proportion of variance that occurs between schools. A value of zero says that schools do not affect the variables and two-level modeling is not needed. More common in studies of school outcomes are ICCs in the range of .10 to .25 (Heck, 2001, p. 99).

Model 4. Controlled composition effects.

The controlled composition effects model adds predictors at both within-school and between-school levels to the baseline two-level model, using the same pair of equations listed on page one of this appendix. The matrices for the within- and between-school models are listed and described separately. The model contains 14 variables. The outcome variable, mrpcm1, is group-mean centered; its within-school mean is fixed at zero ($\nu_1 = \alpha_1 = 0$), while its between-school mean (α_1) is an estimated parameter.

The within-school predictor variables ($y_{2\text{-}8}$) are grand-mean centered; their within-school and between-school means are fixed at zero ($\nu_{2\text{-}8} = \alpha_{2\text{-}8} = 0$). Their between-school variances are fixed at zero for ease of interpretation ($\psi_{22} = \psi_{33} = \psi_{44} = \psi_{55} = \psi_{66} = \psi_{77} = \psi_{88} = 0$). Their within-school variances are fixed at

sample values (see within-school Ψ matrix). Within schools, each predicts the outcome variable (mrpcm1); these are the only relationships modeled in the within-school B-matrix.

The between-school predictor variables ($y_{9\text{-}14}$) are also grand-mean centered ($\nu_{9\text{-}14} = \alpha_{9\text{-}14} = 0$). Their within-school variances are fixed at zero and their between-school variances are fixed at sample values. They each predict between-school adjusted mean values (the between component) of the outcome variable, mrpcm1. These are the only relationships modeled in the between-school B-matrix. All other potential parameters in these models are fixed at zero.

This is essentially a pair of regression models simultaneously estimated. As such, it has zero degrees of freedom (saturated) and is identified. The MPLUS MLR estimator is used. The parameters are maximum-likelihood estimated; the robust standard errors are calculated with a sandwich estimator. No chi-square model fit test statistics are available for saturated models.

Within-school matrices.

y^T = [mrpcm1 title1w black asian hispanic amind other rlunchw rpctfl rpctt1 rpctasn rpctblk rpcthsp rpctind]

ν^T = [0 0 0 0 0 0 0 0 0 0 0 0 0 0]

Λ = diag [1 1 1 1 1 1 1 1 1 1 1 1 1 1]

$\Theta = \mathbf{0}$

α^T = [0 0 0 0 0 0 0 0 0 0 0 0 0 0]

$$B = \begin{bmatrix} 0 & \beta_{12} & \beta_{13} & \beta_{14} & \beta_{15} & \beta_{16} & \beta_{17} & \beta_{18} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\Psi = \begin{bmatrix} \psi_{11} & & & & & & & & & & & & & \\ 0 & .10 & & & & & & & & & & & & \\ 0 & 0 & .07 & & & & & & & & & & & \\ 0 & 0 & 0 & .02 & & & & & & & & & & \\ 0 & 0 & 0 & 0 & .05 & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & .01 & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & .00 & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .10 & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Between-school matrices.

y^T = [mrpcm1 title1w black asian hispanic amind other rlunchw rpctfl rpctt1 rpctasn

rpctblk rpcthsp rpctind]

$$\nu^T = [\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

$$\Lambda = \text{diag}\ [\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ ]$$

$$\Theta = \mathbf{0}$$

$$\alpha^T = [\ \alpha_1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

$$B = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_{19} & \beta_{1,10} & \beta_{1,11} & \beta_{1,12} & \beta_{1,13} & \beta_{1,14} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

$$\Psi = \begin{bmatrix}
\psi_{11} & & & & & & & & & & & & & \\
0 & 0 & & & & & & & & & & & & \\
0 & 0 & 0 & & & & & & & & & & & \\
0 & 0 & 0 & 0 & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .041 & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .063 & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .004 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .034 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .020 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .003
\end{bmatrix}$$

Model 5. **Full-School Engagement Mediation Model**

This final model adds Full-School Engagement (FSE) to the between-schools level of Model 4. FSE is modeled as mediating the relationships of the between-school variables with the between-school intercept of the outcome variable. Computing limitations do not allow for simultaneous modeling of the second-order FSE latent variable and the path model in which it is embedded. For this reason, sum scores are used to define the components of Full-School Engagement, which is then modeled as a latent variable with those components as indicators. The basic equations provided at the beginning of this chapter apply to this model. The matrices involved are described and listed below. Four variables are added to the 14 observed variables of Model 4 – snonres, seng, teng, and peng. Each is created by orienting all indicator variables in the same direction and summing[110]. The Λ-matrices, both between and within-school, become non-square (18 by 15) as the four new observed dependent variables label the four new rows at the top of the matrix and the new latent variable (fse) labels a new column at the beginning of the matrix. A measurement model is added between schools with the estimation of four loadings (λ_1, λ_2, λ_3, and λ_4), four intercepts (ν_1, ν_2, ν_3, and ν_4), and four residual variances (θ_{11}, θ_{22}, θ_{33}, and θ_{44}). The B-matrix becomes 15 by 15 with fse added as the second row and column. Seven parameters are added to the B-matrix in that row. Six are predictors of fse and the seventh measures the effect of fse on mrpcm1. The Ψ-matrix is changed only

[110] One variable, tquit, was also rescaled by a factor of 4/7 to put it on the same four-point scale used by all other observed indicator variables.

by the addition of fse in row and column 1. The variance of fse is fixed at zero. With 23 degrees of freedom, this model satisfies the necessary t-rule for identification, but this is not sufficient. Because of the complexity of the model, it is identified empirically. The MLR estimator, described above, is used.

Within-school matrices.

y^T = [snonres seng teng peng mrpcm1 title1w black asian hispanic amind other rlunchw rpctfl rpctt1 rpctasn rpctblk rpcthsp rpctind]

$$\nu = [\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

$$\Lambda = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

$$\Theta = \mathbf{0}$$

$$\alpha^T = [\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

$$B = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \beta_{32} & \beta_{42} & \beta_{52} & \beta_{62} & \beta_{72} & \beta_{82} & \beta_{92} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

$$\Psi = \begin{bmatrix}
0 & & & & & & & & & & & & & & \\
0 & \psi_{22} & & & & & & & & & & & & & \\
0 & 0 & .099 & & & & & & & & & & & & \\
0 & 0 & 0 & .068 & & & & & & & & & & & \\
0 & 0 & 0 & 0 & .015 & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & .051 & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & .008 & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .003 & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .103 & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

Between-school matrices.

y^T = [snonres seng teng peng mrpcm1 title1w black asian hispanic amind other rlunchw rpctfl rpctt1 rpctasn rpctblk rpcthsp rpctind]

$$\nu^T = [\ \nu_1\ \nu_2\ \nu_3\ \nu_4\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

$$\Lambda = \begin{bmatrix}
\lambda_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\lambda_{21} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\lambda_{31} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\lambda_{41} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

$$\Theta = \begin{bmatrix} \phi_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \phi_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_{33} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \phi_{44} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\alpha^T = [\ 0\ \alpha_2\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_{1,10} & \beta_{1,11} & \beta_{1,12} & \beta_{1,13} & \beta_{1,14} & \beta_{1,15} \\ \beta_{21} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_{2,10} & \beta_{2,11} & \beta_{2,12} & \beta_{2,13} & \beta_{2,14} & \beta_{2,15} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\Psi = \begin{bmatrix}
0 & & & & & & & & & & & & & & \\
0 & \psi_{22} & & & & & & & & & & & & & \\
0 & 0 & 0 & & & & & & & & & & & & \\
0 & 0 & 0 & 0 & & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .041 & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .064 & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .004 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .034 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .021 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .003
\end{bmatrix}$$

APPENDIX B - MEASURES OF FULL-SCHOOL ENGAGEMENT

Two codes are provided for each measure. The leading code, bolded in parenthesis and followed by a colon, is the code used in the MPLUS analyses of this study. The trailing code, in italics and square brackets, is the code provided with the NAEP database.

STUDENT ENGAGEMENT (seng or sdis)

- **% students absent on an average day (sabspct):** About what percentage of your students is absent on an average day? (Include excused and unexcused absences in calculating this rate.) (4 categories) *[C033601]*
- **Students' academic achievement attitudes (sachatt):** How would you characterize each of the following within your school? Students' attitudes toward academic achievement (4 categories) *[C032553]*
- **Student absenteeism problem (sabsprb):** To what degree is each of the following a problem in your school? Student absenteeism (4 categories) *[C032452]*
- **Student tardiness problem (strdprb):** To what degree is each of the following a problem in your school? Student tardiness (4 categories) *[C032451]*

STUDENT RESISTANCE (sres)

- **Physical conflicts among students problem (fightprb):** To what degree is each of the following a problem in your school? Physical conflict among students (4 categories) *[C032454]*
- **Racial or cultural conflicts problem (raceprb):** To what degree is each of the following a problem in your school? Racial or cultural conflicts (4 categories) *[C032457]*

- **Gang activities problem (gangprb):** To what degree is each of the following a problem in your school? Gang activities (4 categories) *[C032463]*
- **Student misbehavior in class problem (smisbprb):** To what degree is each of the following a problem in your school? Student misbehavior in class (4 categories) *[C032464]*
- **Physical conflict between students and teachers problem (stftprb):** To what degree is each of the following a problem in your school? Physical conflicts between students and teachers (4 categories) *[C043153]*
- **Vandalism problem (vandlprb):** To what degree is each of the following a problem in your school? Vandalism (4 categories) *[C043154]*
- **Regard for school property (propreg):** How would you characterize each of the following within your school? Regard for school property (4 categories) *[C032556]*

TEACHER ENGAGEMENT (teng or tdis)

- **Morale of teachers (tmorale):** How would you characterize each of the following within your school? Morale of teachers (4 categories) *[C032552]*
- **Teachers' expectations for student achievement (texpect):** How would you characterize each of the following within your school? Teachers' expectations for student achievement (4 categories) *[C043251]*
- **Percentage of teachers leave before end of year (tquit):** Of the full-time teachers who started in your school last year, what percentage left before the end of the school year? (7 categories) *[C038001]*
- **Teacher absenteeism problem (tabsprb):** To what degree is each of the following a problem in your school? Teacher absenteeism (4 categories) *[C032456]*

- **Percentage of teachers absent on average day (tabspct):** About what percentage of your teachers are absent on an average day? (Include all absences in calculating this rate.) (4 categories) *[C036501]*

PARENT ENGAGEMENT (peng or pdis)

- **Parent support for student achievement (parsupp):** How would you characterize each of the following within your school? Parental support for student achievement (4 categories) *[C032555]*
- **Lack of parent involvement is a problem (pinvprb):** To what degree is each of the following a problem in your school? Lack of parent involvement (4 categories) *[C032459]*
- **Percent involved in making school curriculum decisions (currdec):** In your school, approximately what percentage of the parents do each of the following? Are involved in making school curriculum decisions (4 categories) *[C037704]*
- **Percent at open house or back-to-school nights (opnhouse):** In your school, approximately what percentage of the parents do each of the following? Participate in open house or back-to-school nights (4 categories) *[C037702]*
- **Percent that participate in volunteer programs (volnteer):** In your school, approximately what percentage of the parents do each of the following? Participate in volunteer programs (4 categories) *[C037705]*
- **Percent at parent-teacher conferences (ptconf):** In your school, approximately what percentage of the parents do each of the following? Participate in parent-teacher conferences (4 categories) *[C037703]*

- **Percent that participate in a PTO (pto):** In your school, approximately what percentage of the parents do each of the following? Participate in a parent-teacher organization (4 categories) *[C037701]*

ADMINISTRATIVE OPTIMISM (admpess)

This final latent variable shares measures with the other latent variables because many of those measures are Administrative reports. For example, an administrative report that morale of teachers is a problem in the school may say as much about the optimism of the administrator as about morale of teachers. The following are the eight measures of administrative optimism. The latent variable that shares the measure is in parenthesis.

pinvprb, parsupp (Parent engagement).

tmorale, texpect (Teacher engagement).

sachatt (Student engagement).

gangprb, smisbprb, propreg (Student resistance).

APPENDIX C - 2003 NAEP SAMPLING DESIGN

NAEP uses a stratified, two-stage sampling design. Students are sampled from selected public and nonpublic schools. The sample is stratified on critical subpopulations to ensure adequate representation. Explicit and implicit stratification are used. Within strata, schools are selected with probability proportional to the number of grade-eligible students in the school. Since the passage of No Child Left Behind in 2002, NAEP has combined state public school samples with a national private school sample to produce national estimates. This has increased NAEP's sample size by a factor of nearly 10.

Public schools are explicitly stratified by state. Within each state, schools are hierarchically organized into a series of nested levels.

1. Charter school/non-charter school (two categories)
2. Level of urbanization (eight categories)
3. Percentage of minority students in school (three categories, based on two largest minorities in state)
4. Average achievement of jurisdiction or median income of zip code (continuous)

Approximately 100 schools per state are then systematically sampled with probability proportional to the number of grade-eligible students in the school. The systematic sampling across the hierarchical ordering creates an implicit stratification that tends to enforce proportional representation across the nested levels. In each of the 100 schools, 60 students are randomly sampled – 30 for each of two testing sessions covering two subject areas. Mathematics was one of the subject areas tested in 2003.

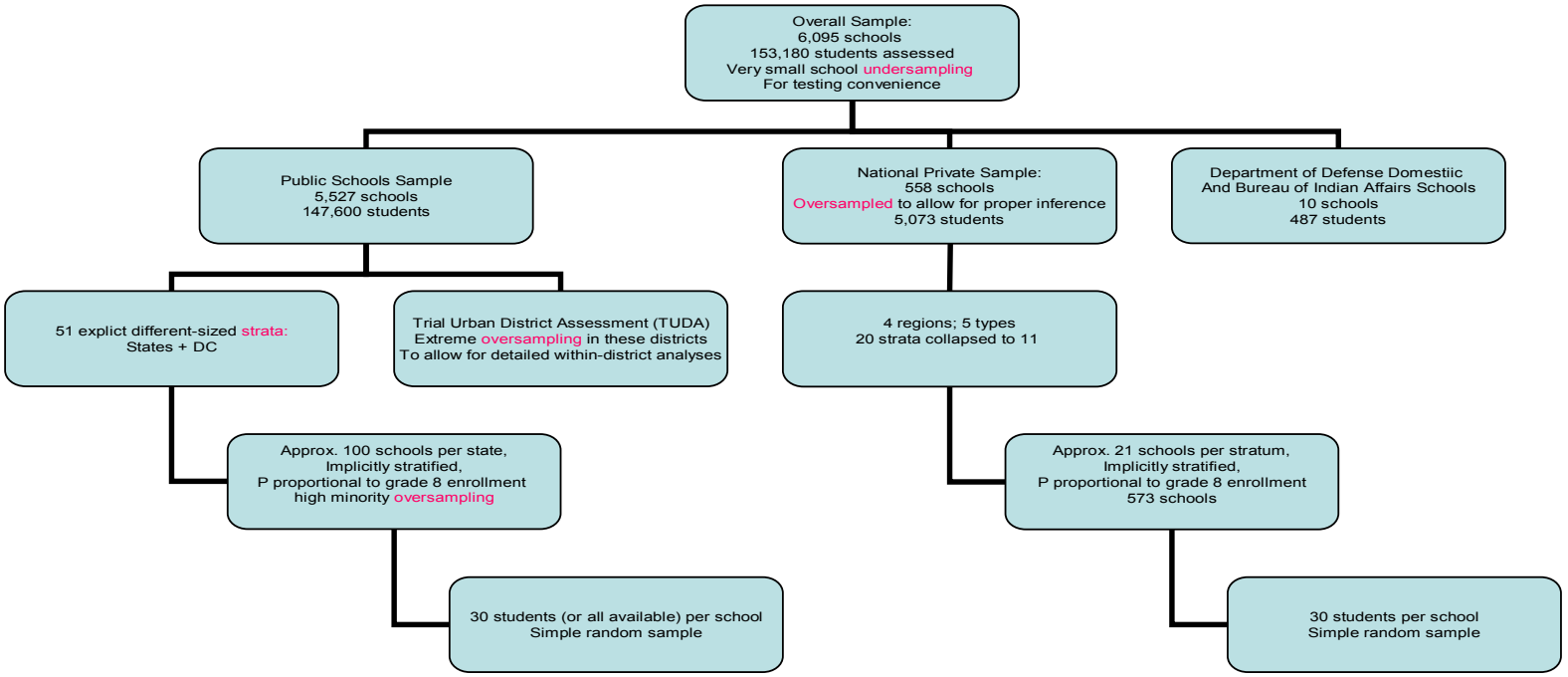
A supplementary national survey of private schools was also conducted. These schools were explicitly stratified into four regions (Northeast, South, Midwest, and West) and five affiliations (Roman Catholic, Lutheran, Conservative Christian, other private, and private type unknown). Implicit stratification (as described for the public school sample) used the following nested hierarchy.

1. Census divisions (nine categories)
2. Urbanization (nine categories)
3. Percentage minority students in school (three categories)

Strata were collapsed as needed to ensure a sufficient number of schools in each stratum. Between 2,500 and 3,000 students were sampled in each state for Mathematics 2003, with an additional 12,600 private school students added to complete the national sample. Overall, about 153,000 students were sampled to represent a target population of 3,938,000 grade 8 students. Weighting variables are provided at both school and student levels.

Participation rates were at least 90% (Braswell et al., 2005, pp. 148-149). Of these students, 10,747 had disabilities or limited English proficiency and were assessed without accommodations, while 11,056 were in these categories but assessed with accommodations (Braswell et al., 2005, pp. 144-145). 5,910 sampled students, a weighted 3% of the target population, were excluded from the testing for reasons of disability or limited English proficiency (Braswell et al., 2005, p. 159).

NAEP 2003 Mathematics Grade 8 Stratified 2-Stage Sample



APPENDIX D - GLOSSARY

adjusted school mean test scores: Many studies have compared school mean test scores. The two-level models in this study allow for a comparison of *adjusted school mean test scores* to try to ensure that the samples being compared have similar characteristics. Each adjusted mean provides the expected test score for a student of nationally average ethnicity and economic level at the given school based on an assumption that every school is equally effective (or ineffective) with all groups of students in the school. This study focuses on mathematics test scores, but the phrase *test scores* is used for brevity. The author expects that study results would be little altered by a focus on language rather than mathematics tests. See also: *grade 8 adjusted school mean mathematics test scores.*

composition effects: This study hypothesizes that school ethnic or economic composition are related to adjusted school mean test scores. These relationships are called *composition effects*. They are sometimes called *compositional effects* or *contextual effects* in the literature.

confirmatory factor analysis (CFA): When a latent variable is defined, one should always statistically analyze the relationships that exist between the proposed variable and its indicators. This process is called a confirmatory factor analysis. There are many ways to analyze the relationships, but the point of each method is to confirm that the factors chosen are appropriate. This study uses a structural equation modeling approach to CFA.

construct: See *latent variable*.

control: Most of the models tested in this study include a number of variables. The regression coefficients obtained by each model are estimated with a *control* for each of

the other variables, meaning that the effect estimates attained are based on the assumption that all other factors are held constant.

covariance matrix: A covariance matrix is the essential data source for a structural equation model. Below the diagonal, the matrix contains the covariances of each pair of observed variables in the study. Above the diagonal, the matrix may either contain the same values (because variable *A*'s covariance with variable *B* is the same as variable *B*'s covariance with variable *A*), or be empty to avoid redundancy. On the diagonal, the matrix contains the covariance of each variable with itself; this is the variance of that variable. For this reason, the covariance matrix is sometimes called a variance/covariance matrix. In some cases, the structural equation modeling software is presented with a covariance matrix; in other cases, it is presented with raw data and calculates the covariance matrix as a first step in model estimation.

database: a collection of datasets. The NAEP Grade 8 Mathematics 2003 database consists of two datasets: a schools dataset and a students dataset. The two are merged in this study to create a third dataset – the merged dataset.

dataset: a collection of quantitative data suitable for statistical analysis. The datasets obtained from NAEP for this study are a *schools dataset* and a *students dataset*. The schools dataset has about 6,000 records, one per school studied. Each record contains a large set of variables; the most substantial are administrative responses to questions about the school. The students dataset contains about 153,000 records. Each record contains information about one of the 20-30 students assessed in a given school. The information includes responses to a student survey, responses to a teacher survey, student responses to a subset of mathematics assessment items, and five plausible values for the

student's mathematics scale score based on those responses as well as responses to the background variables. Many of the models in this study use a *merged dataset*.

economic level: In this study, the economic level of a student is determined by the young person's free and reduced-price lunch status.

effective schools: This study refers to schools that produce higher test scores, even controlling for the backgrounds of the students in the schools, as "effective schools." There is a large body of research, called effective schools research, investigating the characteristics of such schools (American Association of School Administrators, 1992; Cohen et al., 2003; Hawley, 2002; Levine & Lezotte, 1995; Newmann, 1992; Patchen, 2004)

ethnic and economic test score gaps: The U.S. is stratified along both economic and ethnic lines. These two forms of stratification are highly interrelated, but neither is reducible to the other. This stratification is evident in the test scores of students in the nation's schools. Controlling for economic level, White and Asian American students score more highly than Hispanic, Black, and American Indian students. Controlling for ethnicity, more well-to-do students score more highly than less well-to-do students. These differences are called *ethnic and economic test score gaps*.

factor: See *latent variable*.

first-order latent variable: By far, the most common kinds of latent variables are first-order. Like all latent variables, they are measured indirectly. Their values are based on the shared variance of a set of (usually at least three) observed indicator variables.

free or reduced-price lunch status: The most common measure of student economic level in the U.S. is provided by the National School Lunch Program. Lower-income parents

complete applications for *free or reduced-price lunch*. The poorest children receive free lunches; children not as poor receive reduced-price lunches.

grade 8 adjusted school mean mathematics test scores: the most complete signifier for this study's between-schools outcome variable is *grade 8 adjusted school mean mathematics test scores*. It is abbreviated in various ways, for various reasons. For example, "grade 8" is omitted when referring to students at a wider range of grade levels. "Mathematics" is omitted when referring to a broader group of test scores. "School mean" is omitted when referring to individual test scores, as in the within-school model. "Adjusted" is omitted when referring to true school means rather than to means controlled for the ethnicity and economic level of individual students. Sometimes, if the phrase has already been used in a sentence or paragraph, this is abbreviated to "test scores" or "scores."

hierarchical linear model (HLM): See *multi-level model*.

identification: Bollen (1989) provides rules for the identification of structural equation models. In an identified model, a unique mathematical solution exists for each of the estimated structural parameters. In an underidentified model, no unique solution is possible. Identified models can be overidentified or just-identified. In an overidentified model, more information is available than is needed for a solution, allowing for tests of overall model fit. In a just-identified model (such as most OLS regression models), a solution exists, but there is no excess information to allow for tests of model fit.

identified: See *identification*.

indicator: See *latent variable*.

involuntary minorities: John Ogbu uses the concept of i*nvoluntary minorities* to understand ethnic stratification in cross-cultural perspective. Voluntary minority groups are a part of

a nation because of choice – desire for greater opportunity, freedom, etc. Involuntary minority groups have been forcibly incorporated into a nation. The majority group generally wields the greatest power in a nation. Ogbu believes that these categories help explain ethnic test score gaps. In the U.S., the highest test scores are achieved by the majority group (White) and the voluntary minority group (Asian American). The lowest are held by involuntary minorities – a group of conquered peoples (American Indian) and a people brought in slave ships (most Blacks). A middle ground is held by a diverse group of Hispanics. Some, like Asian Americans, are in the U.S. in search of opportunity. Others, like many Chicanos in California, New Mexico, Arizona, and Texas, are members of a conquered community. As a banner carried in a recent Chicano student walkout declared, "We didn't cross the border, the border crossed us."

IRT: Item Response Theory, a statistical method that puts test-takers and test items on the same scale. This allows valid judgments of item difficulty and comparison of scores across tests.

just-identified: See *identification.*

latent variable: like any variable, a statistical representation of an important concept. Unlike most variables, latent variables are not directly observed. Their values are instead inferred from the shared variance of a set of variables presumed to be affected by the latent variable. These variables are called *indicators*, or *measures,* of the latent variable. See also *first-order latent variable, construct,* and *second-order latent variable*. Latent variables are sometimes called factors, or constructs.

listwise deletion: A method of handling missing data in which only records missing no data at all are included in the analysis.

mathematics test scores: See *grade 8 adjusted school mean mathematics test scores, adjusted school mean test scores.*

measure: See *latent variable*.

merged dataset: The merged dataset for this study was created by the addition of a row from the schools dataset to the record of each student who attends the school represented by that row. It has the same number of records as the students dataset (one per targeted student), but the records contain variables from both datasets.

multi-level model: In the U.S., students are clustered within schools, which are clustered within districts, which are clustered within states. Many traditional statistical models focus on students without accounting for this clustering. Two-level, three-level, and even four-level models can improve our understanding of the factors the affect student test scores. See also *single-level model, two-level model.* Multi-level models are sometimes called hierarchical linear models or HLMs.

overidentified: See *identification.*

pairwise deletion: A method of handling missing data in which all records containing complete data for the variables needed for a particular computation are included for that computation. With this method, some elements of a covariance matrix, for example, may be based on more records than other elements.

residual: In structural equation models, observed variables are presumed to be caused by other variables. These causal relationships are designated by arrows leading from cause to effect. The variance in the observed variables is not, however, presumed to be completely explained by the model. Unexplained variance is called the residual and is considered to be the result of (a) variables that are not in the model and (b) random error.

school mean…test scores: See *grade 8 adjusted school mean mathematics test scores, adjusted school mean test scores.*

scores: See *grade 8 adjusted school mean mathematics test scores, adjusted school mean test scores.*

second-order latent variable: Second-order latent variables are much less common in the statistical literature than first-order variables. Like first-order latent variables, their value is based on the shared variance of a set of indicators; they differ in that their indicators include latent variables, not just observed variables.

single-level model: A model in which all important sources of variance are presumed to occur at a single level. An alternative view of this presumption is that there is no significant clustering of subjects within meaningful units. In the U.S. system, this assumption is almost always violated because students are clustered within schools, which are clustered within districts, etc. Meaningful inputs occur at each level. See also *multi-level model, two-level model*.

test scores: See *grade 8 adjusted school mean mathematics test scores*, *adjusted school mean test scores.*

two-level model: A model in which variance is presumed to occur at two levels. Many of the models in this study are two-level. The variance in student test scores is split into within-school and between-school variance. Predictors are used at each level to attempt to explain the variance at that level. For example, student ethnicity is a within-school predictor, whereas school ethnic composition is a between-school predictor. See also *multi-level model, single-level model.*

underidentified: See *identification.*

variance/covariance matrix: See *covariance matrix*.

REFERENCES

Ainsworth-Darnell, J. W., & Downey, D. B. (1998). Assessing the oppositional culture explanation for racial/ethnic differences in school performance. *American Sociological Review, 63*(4), 536-553.

Allington, R. L. (Ed.). (2003). *Big Brother and the National Reading Curriculum*: Heinemann.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545-557.

Alwin, D. F. (1976). Assessing school effects: Some identities. *Sociology of Education, 49*, 294-303.

American Association of School Administrators. (1992). *An effective schools primer*. Arlington, VA: American Association of School Administrators.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (Fifth ed.). Washington, DC: American Psychological Association.

Anderson, S. E. (1997). Worldmath curriculum: Fighting Eurocentrism in mathematics. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 291-306). Albany: SUNY.

Anyon, J. (1981). Social class and school knowledge. *Curriculum Inquiry, 11*(1), 3-42.

Anyon, J. (1995). Race, social class, and educational reform in an inner-city school. *Teachers College Record, 97*, 69-94.

Artiles, A. J., Klingner, J. K., & Tate, W. F. (2006). Representation of minority students in special education: Complicating traditional explanations. *Educational Researcher, 35*(6), 3-5.

Ascher, M., & Ascher, R. (1997). Ethnomathematics. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education*. Albany: SUNY.

Associated Press. (1999). Charlotte schools ordered to stop busing: Integration achieved, judge says. Retrieved May 27, 2005

Bankston III, C., & Caldas, S. J. (1996). Majority African American schools and social injustice: the influence of de facto segregation on academic achievement. *Social Forces, 75*(2), 535-556.

Barton, A. C., Drake, C., Perez, J. G., Louis, K. S., & George, M. (2004). Ecologies of parental engagement in urban education. *Educational Researcher, 33*(4), 3-12.

Barton, P. E. (2003). *Parsing the achievement gap: Baselines for tracking progress* (Policy Information Report). Princeton, NJ: Educational Testing Service.

Bazin, M., & Tamez, M. (2002). *Math and science across cultures: Activities and investigations from the Exploratorium*. San Francisco, CA: Exploratorium.

Berry III, R. Q. (2002). *Voices of African-American male students: A Portrait of successful middle school mathematics students.* Unpublished PhD dissertation, University of North Carolina, Chapel Hill, NC.

Berry III, R. Q. (2004a). The equity principle through the voices of African American males. *Mathematics Teaching in the Middle School, 10*(2), 100-103.

Berry III, R. Q. (2004b). *Voices of successful African American male middle school mathematics students*. Unpublished manuscript.

Betts, J. R., Rueben, K. S., & Danenberg, A. (2000). *Equal resources, equal outcomes? The distribution of school resources and student achievement in California* (report). San Francisco, CA: Public Policy Institute of California.

Bishop, A. J. (2000, October 7-10, 2000). *Critical challenges in researching cultural issues in mathematics learning.* Paper presented at the Proceedings of the Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Tucson, AZ.

Blau, J. R. (2003). *Race in the schools: Perpetuating white dominance?* Boulder, CO: Lynne Rienner, Publishers.

Boger, J. C., & Orfield, G. (Eds.). (2005). *School resegregation: Must the South turn back?* Chapel Hill: University of North Carolina Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley and Sons.

Bollen, K. A., Glanville, J., & Stecklov, G. (2001). Socioeconomic status and class in studies of fertility and health in developing countries. *Annual Review of Sociology, 27*, 153-185.

Borkan, B., Capa, Y., Figueiredo, C., & Loadman, W. E. (2003, October 15-18, 2003). *Using rasch measurement to evaluate the organizational climate index.* Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Columbus, OH.

Borman, G. D. (2005). National efforts to bring reform to scale in high-poverty schools: Outcomes and implications. *Review of Research in Education, 29*(special issue), 1-27.

Borman, G. D., & D'Agostino, J. V. (1996). Title I and student achievement: A meta-analysis of federal evaluation results. *Educational Evaluation and Policy Analysis, 18*(4), 309-326.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125-230.

Borman, K. M., Eitle, T. M., Michael, D., Eitle, D. J., Lee, R., Johnson, L., et al. (2004). Accountability in a postdesegregation era: The continuing significance of racial segregation in Florida's schools. *American Educational Research Journal, 41*(3), 605-631.

Bourdieu, P. (1973). Cultural reproduction and social reproduction. In R. Arum & I. R. Beattie (Eds.), *The structure of schooling: Readings in the sociology of education* (pp. 56-69). Mountain View, CA: Mayfield Publishing Company.

Bourdieu, P. (1986). The forms of capital (R. Nice, Trans.). In J. G. Richardson (Ed.), *Handbook for theory and research in the sociology of education* (pp. 241-258). New York: Greenwood Press.

Bourdieu, P., & Passeron, J.-C. (1977). *Reproduction in education, society, and culture* (R. Nice, Trans.). Beverly Hills, CA: Sage.

Bourdieu, P., & Passeron, J.-C. (1990). *Reproduction in education, society, and culture.* (R. Nice, Trans. 1990 ed.). London: Sage.

Braswell, J. S., Dion, G. S., Daane, M. C., & Jin, Y. (2005). *The nation's report card: Mathematics 2003* (NCES report No. NCES 2005-451). Washington, DC: U.S. Government Printing Office.

Braun, H., Jenkins, F., & Grigg, W. (2006). *Comparing private schools and public schools using hierarchical linear modeling* (report No. NCES 2006-461). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences.

Brewster, A. B., & Bowen, G. L. (2004). Teacher support and the school engagement of latino middle and high school students at risk of school failure. *Child and Adolescent Social Work Journal, 21*(1).

Bryk, A. S., & Driscoll, M. E. (1988). *The high school as community: Contextual influences and consequences for students and teachers* (report). Madison, WI: National Center on Effective Secondary Schools.

Buckley, M. A., Storino, M., & Sebastiani, A. M. (2003). *The impact of school climate: Variation by ethnicity and gender.* Paper presented at the Annual Conference of the American Psychological Association, Toronto, ON.

Burtless, G. (Ed.). (1996). *Does money matter? The effects of school resources on student achievement and adult success*. Washington, DC: Brookings Institution Press.

Cahalan, M. W., Ingels, S. J., Burns, L. J., Planty, M., & Daniel, B. (2006). *United States high school sophomores: A twenty-two year comparison, 1980-2002* (report No. NCES 2006-327). Washington, DC: National Center for Education Statistics.

Card, D., & Rothstein, J. (2006). *Racial segregation and the black-white test score gap* (working paper). Cambridge, MA: National Bureau of Economic Research.

Cashin, S. (2004). *The failures of integration: How race and class are undermining the American dream*. New York: Public Affairs.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*(2), 294-343.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119-142.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Wienfeld, F. D., et al. (1966). *Equality of Educational Opportunity*. Washington, DC: US Government Printing Office.

Comer, J. P. (2001). Schools that develop children. *The American Prospect, 12*(7), 8.

Comer, J. P. (2004). *Leave no child behind: Preparing today's youth for tomorrow's world*. New Haven: Yale University Press.

Comer, J. P., Haynes, N. M., Joyner, E. T., & Ben-Avie, M. (Eds.). (1996). *Rallying the whole village: The Comer process for reforming education* (1996 ed.). New York: Teachers College Press.

Comer, J. P., Michael, B.-A., Haynes, N. M., & Joyner, E. T. (Eds.). (1999). *Child by child: The Comer process for change in education*. New York, NY: Teachers College Press.

Cook, P. J., & Ludwig, J. (1998). The burden of 'acting White': Do Black adolescents disparage academic achievement? In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 375-400). Washington, DC: Brookings Institution Press.

Crain, R. I., & Mahard, R. E. (1978). Desegregation and black achievement: A review of the research. *Law and Contemporary Problems, 23*(3), 17-56.

D'Ambrosio, U. (1997). Ethnomathematics and its place in the history and pedagogy of mathematics. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 13-24). Albany: SUNY.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1).

Darling-Hammond, L. (2002-2003). Access to quality teaching: An analysis of inequality in California's public schools. *Santa Clara Law Review, 43*, 1045-1184.

Darling-Hammond, L. (2004). The color line in American education: Race, resources, and student achievement. *Du Bois Review, 1*(2), 213-246.

Darling-Hammond, L. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives, 13*(42), 51.

Darling-Hammond, L. (2006). Securing the right to learn: Policy and practice for powerful teaching and learning. *Educational Researcher, 35*(7), 13-24.

Delpit, L. D. (1995). *Other people's children*. New York: New Press.

Delpit, L. D. (2003). Educators as 'seed people' growing a new future. *Educational Researcher, 32*(7), 14-21.

Dewey, J., Husted, T. A., & Kenney, L. W. (2000). The ineffectiveness of school inputs: A product of misspecification? *Economics of education review, 19*, 27-45.

Dinkes, R., Cataldi, E. F., Kena, G., & Baum, K. (2006). *Indicators of school crime and safety: 2006* (report No. NCES 2007-003). Washington, DC: U.S. Departments of Education and Justice.

Driessen, G. W. J. M. (2001). Ethnicity, forms of capital, and educational achievement. *International Review of Education, 47*(6), 513-538.

Education Trust - West. (2005). *California's hidden teacher spending gap: How state and district budgeting practices shortchange poor and minority students and their schools* (report). Oakland, CA: Author.

Education Trust. (2006). *Funding Gaps 2006* (report). Washington, DC: Author.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1-26.

Esch, C. E., Chang-Ross, C. M., Guha, R., Humphrey, D. C., Shields, P. M., Tiffany-Morales, J. D., et al. (2005). *Teaching and California's future: The status of the teaching profession 2005* (report). Santa Cruz, CA: The Center for the Future of Teaching and Learning.

Evans, W. N., Murray, S. E., & Schwab, R. M. (1997). Schoolhouses, courthouses, and statehouses after *Serrano*. *Journal of Policy Analysis and Management, 16*(1), 10-31.

Everson, H. T., & Millsap, R. E. (2004). *Beyond individual differences: Exploring school effects on SAT scores* (report No. 2004-3). New York: College Entrance Examination Board.

Fasheh, M. (1997). Mathematics, culture, and authority. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 273-289). Albany: SUNY.

Ferguson, A. A. (2000). *bad boys: Public Schools in the Making of Black Masculinity*. Ann Arbor: The University of Michigan Press.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28*, 465-498.

Ferguson, R. F. (1998). Teachers' perceptions and expectations and the Black-White test score gap. In C. Jencks & M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 318-374). Washington, DC: Brookings Institution Press.

Figlio, D. N. (1999). Functional form and the estimated effects of school resources. *Economics of education review, 18*, 241-252.

Finn, J. D., & Voelkl, K. E. (1993). School characteristics related to school engagement. *The Journal of Negro Education, 62*(3), 249-268.

Flinspach, S. L., & Banks, K. E. (2005). Moving beyond race: Socioeconomic diversity as a race-neutral approach to desegregation in the Wake County Schools. In J. C. Boger & G. Orfield (Eds.), *School resegregation: Must the South turn back?* (pp. 261-280). Chapel Hill: University of North Carolina Press.

Fordham, S., & Ogbu, J. U. (1986). Black students' school success: Coping with the burden of 'Acting White'. *The Urban Review, 18*(3), 176-206.

Freeman, C. E., Scafidi, B., & Sjoquist, D. L. (2005). Racial segregation in Georgia public schools, 1994-2001. In J. C. Boger & G. Orfield (Eds.), *School resegregation: Must the South turn back?* (pp. 148-163). Chapel Hill: University of North Carolina Press.

Gamoran, A. (1992). The variable effects of high school tracking. *American Sociological Review, 57*(6), 812-828.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of teacher education, 53*(2), 106-116.

Gerdes, P. (1997a). On culture, geometrical thinking and mathematics education. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 223-247). Albany: SUNY.

Gerdes, P. (1997b). Survey of current work on ethnomathematics. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 331-371). Albany: SUNY.

Gingerich, D. (2003). No Child Left Behind. *Currents, 6*(2), 1, 12-14.

Ginsburg, H. P. (1986). The myth of the deprived child: New thoughts on poor children. In U. Neisser (Ed.), *The school achievement of minority children: New perspectives* (pp. 169-189). Hillsdale, NJ: Erlbaum.

Goe, L. (2002). Legislating equity: The distribution of emergency permit teachers in California. *Education Policy Analysis Archives, 10*(42).

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter?: Assessing the impact of unobservables on educational productivity. *The journal of human resources, 32*(3), 505-520.

Green, P. J., Dugoni, B. L., Ingels, S. J., & Camburn, E. (1995). *A profile of the American high school senior in 1992*. Washington, DC: U.S. Department of Education.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361-396.

Gregoire, M., & Algina, J. (2000, April 24-28, 2000). *Reconceptualizing the debate on school climate and students' academic motivation and achievement: A multilevel analysis.* Paper presented at the American Educational Research Association, New Orleans, LA.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: RAND.

Grissmer, D., Flanagan, A., & Williamson, S. (1998). Why did the black-white score gap narrow in the 1970s and 1980s? In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 182-226). Washington, DC: Brookings Institute.

Guerino, P., Hurwitz, M. D., Noonan, M. E., & Kaffenberger, S. M. (2006). *Crime, violence, discipline, and safety in U.S. public schools: Findings from the school survey on crime and safety: 2003-04* (report No. nces 2007-302). Washington, DC: National Center for Education Statistics.

Gutiérrez, R. (2002). Enabling the practice of mathematics teachers in context: Toward a new equity research agenda. *Mathematical Thinking and Learning, 4*(2&3), 145-187.

Gutstein, E., Lipman, P., Hernandez, P., & Reyes, R. d. l. (1997). Culturally Relevant Mathematics Teaching in a Mexican American Context. *Journal for Research in Mathematics Education, 28*(6), 709-737.

Hannah-Jones, N. (2006, March 15, 2006). Best teachers not where needed. *News and Observer,* pp. 1, 10.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 29*, 1141-1177.

Hanushek, E. A. (1996a). A more complete picture of school resource policies. *Review of Educational Research, 66*(3), 397-409.

Hanushek, E. A. (1996b). School resources and student performance. In G. Burtless (Ed.), *Does money matter? The effects of school resources on student achievement and adult success* (pp. 43-73). Washington, DC: Brookings Institution Press.

Harris, L. (2004). *Report on the status of public school education in California 2004* (report). Los Angeles: UCLA Institute for Democracy, Education, and Access.

Harris, M. (1997). An example of traditional women's work as a mathematics resource. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 215-222). Albany: SUNY.

Harris, M. M., & Willomer, D. J. (1998). Principal's optimism and perceived school effectiveness. *Journal of Educational Administration, 36*, 353-361.

Harry, B., & Klingner, J. (2006). *Why are so many minority students in special education?* New York: Teachers College Press.

Hawley, W. D. (Ed.). (2002). *The keys to effective schools: Education reform as continuous improvement*. Thousand Oaks, CA: Corwin Press, Inc.

Heath, S. B. (1982). Questioning at home and at school: A comparative study. In G. D. Spindler (Ed.), *Doing the ethnography of schooling*. New York: Holt, Rinehart & Winston.

Heck, R. H. (2001). Multilevel modeling with SEM. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 89-127). Mahwah, NJ: Lawrence Erlbaum Associates.

Hedges, L. V., & Greenwald, R. (1996). Have times changed? The relation between school resources and student performance. In G. Burtless (Ed.), *Does money matter? The effects of school resources on student achievement and adult success* (pp. 74-92). Washington, DC: Brookings Institution Press.

hooks, b. (1994). *Teaching to transgress: Education as the practice of freedom*. New York: Routledge.

Hoover-Dempsey, K. V., & Sandler, H. M. (1997). Why do parents become involved in their children's education? *Review of Educational Research, 67*(1), 3-42.

Horkay, N. (Ed.). (1999). *The NAEP guide*. Washington, DC: National Center for Education Statistics.

Howard, G. R. (1999). *We can't teach what we don't know: White teachers, multiracial schools*. New York: Teachers College Press.

Jefferson, A. L. (2005). Student performance: Is more money the answer? *Journal of Education Finance, 31*(2), 111-124.

Jencks, C., & Phillips, M. (Eds.). (1998). *The black-white test score gap*. Washington, DC: Brookings Institution Press.

Jenkins, P. H. (1995). School delinquency and school commitment. *Sociology of Education, 68*(3), 221-239.

Johnson, K. A. (2000). *Comparing math scores of black students in D.C.'s public and catholic schools*. Washington, DC: Heritage Foundation, Center for Data Analysis. (ERIC Document Reproduction Service No. ED440209).

Johnson, M. K., Crosnoe, R., & Elder Jr., G. H. (2001). Students' attachment and academic engagement: The role of race and ethnicity. *Sociology of Education, 74*(4), 318-340.

Jöreskog, K. (1999, June 22, 1999). How large can a standardized coefficient be? Retrieved May 25, 2007, from http://www.ssicentral.com/lisrel/techdocs/HowLargeCanaStandardizedCoefficientbe.pdf

Joseph, G. G. (1997). Foundations of Eurocentrism in mathematics. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 61-81). Albany: SUNY.

Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 34*(8), 3-13.

King, L. (2006, April 26, 2006). Gap in teacher quality falls on income lines. *USA Today,* p. 2.

Kochman, T. (1981). *Black and white styles in conflict*. Chicago: The University of Chicago Press.

Kozol, J. (1991). *Savage Inequalities: Children in America's Schools*. New York: Crown Publishers, Inc.

Kozol, J. (2005). *The shame of the nation: The restoration of apartheid schooling in America*. New York: Crown Publishers.

Krei, M. S. (2000, April 2000). *Teacher transfer policy and the implications for equity in urban school districts.* Paper presented at the Annual Meeting of the American Educational Researchers Association, New Orleans.

Ladson-Billings, G. (1994). *The Dreamkeepers: Successful Teachers of African American Children*. San Francisco: Jossey-Bass Publishers.

Ladson-Billings, G. (1997). It doesn't add up: African American students' mathematics achievement. *Journal for Research in Mathematics Education, 28*(6), 697-708.

Ladson-Billings, G. (2001). The power of pedagogy: Does teaching matter? In W. H. Watkins, J. H. Lewis & V. Chou (Eds.), *Race and education: The roles of history and society in educating African American students* (pp. 73-88). Boston: Allyn and Bacon.

Ladson-Billings, G. (2004). Landing on the wrong note: The price we paid for *Brown*. *Educational Researcher, 33*(7), 3-13.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37-62.

Lee, J.-S., & Bowen, N. K. (2006). Parent involvement, cultural capital, and the achievement gap among elementary school children. *American Educational Research Journal, 43*(2), 193-218.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31*(1), 3-12.

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal, 41*(4), 797-832.

Lee, V. E., & Bryk, A. S. (1988). Curriculum tracking as mediating the social distribution of high school achievement. *Sociology of Education, 61*, 78-94.

Lee, V. E., & Bryk, A. S. (1989). a multilevel model of the social distribution of high school achievement. *Sociology of Education, 62*, 172-192.

Lee, V. E., Bryk, A. S., & Smith, J. B. (1993). The organization of effective secondary schools. In L. Darling-Hammond (Ed.), *Review of Research in Education* (Vol. 19, pp. 171-267). Washington, DC: American Educational Research Association.

Lee, V. E., Dedrick, R. F., & Smith, J. B. (1991). The effect of the social organization of schools on teachers' efficacy and satisfaction. *Sociology of Education, 64*(3), 190-208.

Lee, V. E., & Smith, J. B. (1995). Effects of high school restructuring and size on early gains in achievement and engagement. *Sociology of Education, 68*(4), 241-270.

Levačić, R., & Vignoles, A. (2002). Researching the links between school resources and student outcomes in the UK: A review of issues and evidence. *Education Economics, 10*(3), 313-331.

Levine, D. U., & Lezotte, L. W. (1995). Effective schools research. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of Research on Multicultural Education* (pp. 525-547). New York: Macmillan.

Lindquist, M. M. (2001). NAEP, TIMSS, and PSSM: Entangled Influences. *School Science and Mathematics, 101*(6), 286-291.

Lipka, J., Mohatt, G. V., & Ciulistet Group. (1998). Expanding curricular and pedagogical possiblities: Yup'ik-based mathematics, science, and literacy. In *Transforming the culture of schools: Yup'ik Eskimo examples* (pp. 139-181). Mahwah, NJ: Erlbaum.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (second edition ed.). Hoboken, NJ: Wiley - Interscience.

Louis, K. S., & Smith, B. (1992). Cultivating teacher engagment: Breaking the iron law of social class. In F. M. Newmann (Ed.), *Student engagement and achievement in American secondary schools* (pp. 119-152). New York: Teachers College Press.

Loveless, T. (1999). *The tracking wars: state reform meets school policy*. Washington, DC: Brookings Institution Press.

Lubienski, C., & Lubienski, S. T. (2006). *Charter, private, public schools and academic achievement: New evidence from NAEP mathematics data* (report). New York: National Center for the Study of Privatization in Education.

Lubienski, S. T. (2000). A clash of social class cultures? Students' experiences in a discussion-intensive seventh-grade mathematics classroom. *The Elementary School Journal, 100*(4), 377-403.

Lubienski, S. T. (2001). *A second look at mathematics achievement gaps: Intersections of race, class, and gender in NAEP data.* Paper presented at the Annual meeting of the American Educational Research Association, Seattle, WA.

Lubienski, S. T. (2002). *Are we achieving 'Mathematical power for all?' A Decade of national data on instruction and achievement.* Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, LA.

Lubienski, S. T. (2006). Examining instruction, achievement, and equity with NAEP mathematics data. *Educational Policy Analysis Archives, 14*(14), 33.

Lubienski, S. T., & Bowen, A. (2000). Who's counting? A survey of mathematics education research. *Journal for Research in Mathematics Education, 31*(5), 626-633.

Lubienski, S. T., & Shelley, M. C., II. (2003). *A closer look at U.S. mathematics instruction and achievement: Examinations of race and SES in a decade of NAEP data.* Paper presented at the American Educational Research Association, Chicago.

MacLeod, J. (1995). Teenagers in Clarendon Heights: The Hallway Hangers and the Brothers. In *Ain't No Makin' It: Aspirations and attainment in a low-income neighborhood*: Westview Press.

Malloy, C., & Malloy, W. (1998). Issues of culture in mathematics teaching and learning. *The urban review, 30*, 245-257.

Martin, D. B. (2000). *Mathematics success and failure among African-American youth: The roles of sociohistorical context, community forces, school influence, and individual agency*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29-46.

McCabe, M. (2006a, January 5, 2006). A decade of effort. *Education Week, 25,* 8-21.

McCabe, M. (2006b, January 5, 2006). State of the states. *Education Week, 25,* 86-98.

McCourt, F. (2005). *Teacher man*. New York: Scribner.

Mickelson, R. A. (1990). The attitude-achievement paradox among black adolescents. *Sociology of Education, 63*(1), 44-61.

Miller, G. E. (2003). Analyzing the minority gap in achievement scores: Issues for states and federal government. *Educational Measurement: Issues and Practice, 22*(3), 30-36.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Moses, R. P., & Cobb, C. E., Jr. (2001). *Radical equations: Math literacy and civil rights*. Boston: Beacon Press.

Mullis, I. V. S., Jenkins, F., & Johnson, E. G. (1994). *Effective Schools in Mathematics: Perspectives from the NAEP 1992 Assessment*. Washington, DC: National Center for Education Statistics.

Muthén, B. O. (1998-2004). Mplus technical appendices. 2005, from http://www.statmodel.com/download/techappen.pdf

Muthén, L. K., & Muthén, B. O. (1998-2005). *Mplus User's Guide* (Third ed.). Los Angeles: Muthén and Muthén.

Myers, D. E. (1985). The relationship between school poverty concentration and students' reading and math achievement and learning. In M. M. Kennedy, R. K. Jung & M. E. Orland (Eds.), *Poverty, achievement, and the distribution of compensatory education services* (pp. D15 - D60). Washington, DC: Government Printing Office.

National Center for Education Statistics. (2003). *The nation's report card: Mathematics highlights 2003* (report). Jessup, MD: National Center for Education Statistics.

National Center for Education Statistics. (2004). NAEP Data Tool. Retrieved November 4, 2004, from http://nces.ed.gov/nationsreportcard/naepdata

National Center for Education Statistics. (2005). 2005 Assessment Results: The Nation's Report Card. Retrieved January 9, 2006, from http://nces.ed.gov/nationsreportcard/nrc/reading_math_2005/

National Center for Education Statistics. (2006). *School and parent interaction by household language and poverty status: 2002-03* (Issue Brief No. NCES 2006-086). Washington, DC: National Center for Education Statistics.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Newmann, F. M. (Ed.). (1992). *Student engagement and achievement in American secondary schools*. New York: Teachers College Press.

Nind, M., Rix, J., Sheehy, K., & Simmons, K. (Eds.). (2005). *Curriculum and pedagogy in inclusive education*. New York: RoutledgeFalmer.

Noblit, G. W., Malloy, W. W., & Malloy, C. E. (Eds.). (2001). *"The kids got smarter": Case studies of successful Comer Schools*. Cresskill, NJ: Hampton Press.

Noblit, G. W., & Patterson, J. A. (2001). The school development program and education reform. In G. W. Noblit, W. W. Malloy & C. E. Malloy (Eds.), *"The kids got smarter": Case studies of successful Comer Schools* (pp. 1-16). Cresskill, NJ: Hampton Press.

O'Connor, C. (1998). Resilience despite reproductive notions of risk: A case of black inner-city youth. In K. K. Wong (Ed.), *Advances in educational policy: Perspectives on the social functions of schools* (Vol. 4, pp. 51-86). Stamford, Connecticut: Jai Press, Inc.

Oakes, J., Johnson, R., & Muir, K. (2004). Access and achievement in mathematics and science: Inequalities that endure and change. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of Research on Multicultural Education* (second ed., pp. 69-90). San Francisco: Wiley and Sons.

Oakes, J., Rogers, J., Silver, D., & Goode, J. (2004). *Separate and unequal 50 years after Brown: California's racial 'opportunity gap'* (report). Los Angeles: UCLA/IDEA.

Ogbu, J. U. (1978). *Minority education and caste: The American system in cross-cultural perspective*. New York: Academic Press.

Ogbu, J. U. (1988). Class stratification, racial stratification, and schooling. In L. Weis (Ed.), *Class, Race, and Gender in American Education* (pp. 163-179). Albany: State University of New York.

Ogbu, J. U. (1992). Understanding cultural diversity and learning. *Educational Researcher, 21*(8), 5-14.

Ogbu, J. U. (1997). African American education: A cultural -ecological perspective. In H. P. McAdoo (Ed.), *Black Families* (3rd ed., pp. 234-250). Thousand Oaks, CA: Sage.

Ogbu, J. U., & Simons, H. D. (1994). *Cultural models of school achievement: A quantitative test of Ogbu's theory. Cultural models of literacy: A comparative study. Project 12.*

(report No. CS 214 649). Berkeley, CA: National Center for the Study of Writing and Literacy.

Olson, L. (Ed.). (2007). *Quality counts 2007: From cradle to career* (Vol. 26 (17)). Bethesda, MD: Editorial Projects in Education.

Orfield, G. (1996). The growth of segregation: African Americans, Latinos, and unequal education. In H. Hill & J. James E. Jones (Eds.), *Dismantling desegregation: the quiet reversal of Brown v. Board of Education* (pp. 234-262). Madison, WI: University of Wisconsin Press.

Orfield, G. (2001). Schools more separate. *Rethinking schools online, 16*(1), 10.

Orfield, G., & DeBray, E. H. (Eds.). (1999). *Hard work for good schools: Facts not fads in Title I reform*. Cambridge, MA: The Civil Rights Project.

Orfield, G., & Yun, J. T. (1999). *Resegregation in American schools* (report). Cambridge, MA: The Civil Rights Project, Harvard University.

Patchen, M. (2004). *Making our schools more effective: What matters and what works*. Springfield, IL: Charles C. Thomas.

Perie, M., Grigg, W., & Dion, G. (2005). *The Nation's Report Card: Mathematics 2005* (report No. NCES 2006-453). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Perry, T., Steele, C., & Hilliard, A. (Eds.). (2003). *Young, gifted, and black: Promoting high achievement among African American students*. Boston: Beacon.

Perry, Y., Cartron, K., Gerlach, D., Kingsberry-Burt, S., Dozier, N., Lazo-Chadderton, M., et al. (2000). *Exposing the gap: Why minority students are being left behind in North Carolina's educational system* (report). Raleigh, NC: North Carolina Justice and Community Development Center and the North Carolina Education and Law Project.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525-556.

Phillips, M., Brooks-Gunn, J., Duncan, G. J., Klebanov, P., & Crane, J. (1998). Family background, parenting practices, and the black-white test score gap. In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 103-148). Washington, DC: The Brookings Institution.

Pinxten, R. (1997). An ethnomathematical approach in mathematical education: A matter of political power. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 403-418). Albany: SUNY.

Popkewitz, T. (2004). School subjects, the politics of knowledge, and the projects of intellectuals in charge. In P. Valero & R. Zevenbergen (Eds.), *Researching the socio-political dimensions of mathematics education: Issues of power in theory and methodology* (pp. 251-267). New York: Kluwer.

Porter, M. K. (1996, October 10-15, 1996). *Moving mountains: Reform, resistance and resiliency in an Appalachian Kentucky high school.* Paper presented at the Annual Convention of the National Rural Education Association, San Antonio, TX.

Powell, A. B., & Frankenstein, M. (1997a). Considering interactions between culture and mathematical knowledge. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 119-127). Albany: SUNY.

Powell, A. B., & Frankenstein, M. (1997b). Ethnomathematical praxis in the curriculum. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 249-259). Albany: SUNY.

Powell, A. B., & Frankenstein, M. (1997c). Ethnomathematical research. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 321-330). Albany: SUNY.

Powell, A. B., & Frankenstein, M. (1997d). Ethnomathematics. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 1-3). Albany: SUNY.

Powell, A. B., & Frankenstein, M. (1997e). Reconsidering what counts as mathematical knowledge. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 193-199). Albany: SUNY.

Powell, A. B., & Frankenstein, M. (1997f). Uncovering the distorted and hidden history of mathematical knowledge. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 51-59). Albany: SUNY.

Prediger, S. (2004). Intercultural perspectives on mathematics learning -- Developing a theoretical framework. *International Journal of Science and Mathematics Education, 2*, 377-406.

Pritchett, L., & Filmer, D. (1999). What education production functions *really* show: a positive theory of education expenditures. *Economics of education review, 18*, 223-239.

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2000). HLM 5: Hierarchical linear and nonlinear modeling (Version 5) [MLM software]. Chicago: Scientific Softward International.

Raudenbush, S. W., Fotiu, R. P., & Cheong, Y. F. (1998). Inequality of access to educational resources: A national report card for eighth-grade math. *Educational Evaluation and Policy Analysis, 20*(4), 253-267.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics, 16*(4), 295-330.

Reardon, S. (2005, May 30, 2006). Re: Losing significance by increasing standard errors. Retrieved January 24, 2005, 2005, from http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind05&L=MULTILEVEL&D=0&I=-3&P=11792

Redding, S. (2005). Rallying the troops. *The School Community Journal, 15*(1), 7-13.

Regan, R. (2005). *The relationship between school socioeconomic composition and academic performance: A comparative analysis of elementary schools in the five largest North Carolina school districts 2002-2003.* Paper presented at the Annual Meeting of the North Carolina Association for Research in Education, Chapel Hill, NC.

Rist, R. C. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review, 40*(3), 411-451.

Rogers, A. M., & Stoeckel, J. J. (2004). *NAEP 2003 mathematics and reading assessments secondary-use data files data companion*. Washington, DC: National Center for Education Statistics.

Rogoff, B. (2003). *The cultural nature of human development*. New York: Oxford University Press.

Rogoff, B., & Chavajay, P. (1995). What's become of research on the cultural basis of cognitive development? *American Psychologist, 30*(10), 859-877.

Rosenbaum, J. E. (1995). Changing the geography of opportunity by expanding residential choice: Lessons from the Gautreaux program. *Housing Policy Debate, 6*(1), 231-269.

Ross, S. M., & Lowther, D. L. (2003). Impacts of Co-nect school reform design on classroom instruction, school climate, and student achievement in inner-city schools. *Journal of Education for Students Placed At Risk, 8*(2), 215-246.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the black-white achievement gap*. Washington, DC: Economic Policy Institute.

Rumberger, R. W., & Palardy, G. J. (2005). Does resegregation matter? In J. C. Boger & G. Orfield (Eds.), *School resegregation: Must the South turn back?* (pp. 127-147). Chapel Hill: University of North Carolina Press.

Sable, J., & Hill, J. (2006). *Overview of public elementary and secondary students, staff, schools, school districts, revenues, and expenditures: School year 2004-05 and fiscal year 2004* (report No. nces 2007-309). Washington, DC: National Center for Education Statistics.

Schellenberg, S. J. (1999). Concentration of poverty and the ongoing need for Title I. In G. Orfield & E. H. DeBray (Eds.), *Hard work for good schools: Facts not fads in Title I reform* (pp. 130-146). Cambridge, MA: The Civil Rights Project.

Secada, W. G. (1992). Race, ethnicity, social class, language, and achievement in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning*. New York: MacMillan.

Shindler, J., Taylor, C., Cadenas, H., & Jones, A. (2003, April 21-25, 2003). *Sharing the data along with the responsibility: Examining an analytic scale-based model for assessing school climate.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Silins, H., & Mulford, B. (2004). Schools as learning organisations -- effects on teacher leadership and student outcomes. *School Effectiveness and School Improvement, 15*(3-4), 443-466.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

Sleeter, C. E. (2005). *Un-standardizing curriculum*. New York: Teachers College Press.

Solomon, R. P. (1988). Black cultural forms in schools: A cross national comparison. In L. Weis (Ed.), *Class, Race, and Gender in American Education* (pp. 249-265). Albany: State University of New York.

Spade, J. Z., Columbia, L., & Vanfossen, B. E. (1997). Tracking in mathematics and science: courses and course selection procedures. *Sociology of Education*.

Spring, J. (2004). *Deculturalization and the struggle for equality* (fourth ed.). New York: McGraw Hill.

Stiff, L. V. (1990). African-American students and the promise of the *Curriculum and Evaluation Standards*. In T. J. Cooney & C. R. Hirsch (Eds.), *Teaching and learning mathematics in the 1990s: 1990 Yearbook* (pp. 152-158). Reston, Virginia: The National Council of Teachers of Mathematics.

Street, P. (2005). *Segregated schools*. New York: Routledge.

Strutchens, M. E. (2000). Confronting the beliefs and stereotypes that impede the mathematical empowerment of African American students. In M. E. Strutchens, M. L. Johnson & w. F. Tate (Eds.), *Perspectives on African-Americans*. Reston, VA: National Council of Teachers of Mathematics.

Strutchens, M. E., Lubienski, S. T., McGraw, R., & Westbrook, S. K. (2004). NAEP findings regarding race and ethnicity: Students' performance, school experiences, attitudes and beliefs, and family influences. In P. Kloosterman & J. Frank K. Lester (Eds.), *Results and Interpretations of the 1990-2000 Mathematics Assessments of the National Assessment of Educational Progress* (pp. 269-304). Reston, VA: National Council of Teachers of Mathematics.

Sullivan, A. (2001). Cultural capital and educational attainment. *Sociology, 35*(4), 318-340.

Tate, W. F. (1995). Returning to the root: A culturally relevant approach to mathematics pedagogy. *Theory Into Practice, 34*(3).

Tate, W. F. (1997). Race-ethnicity, SES, gender, and language proficiency trends in mathematics achievement: an update. *Journal for Research in Mathematics Education, 28*, 652-679.

Teranishi, R., Allen, W. R., & Solórzano, D. G. (2004). Opportunity at the crossroads: Racial inequality, school segregation, and higher education in California. *Teachers College Record, 106*(11), 2224-2245.

Tucker, C. M., Zayco, R. A., Herman, K. C., Reinke, W. M., Trujillo, M., Carraway, K., et al. (2002). Teacher and child variables as predictors of academic engagement among low-income African American Children. *Psychology in the Schools, 39*(4), 477-488.

Tyson, K. (2002). Weighing in: Elementary-age students and the debate on attitudes toward school among Black students. *Social Forces, 80*(4), 1157-1189.

Tyson, K. (2003). Notes from the back of the room: Problems and paradoxes in the schooling of young black students. *Sociology of Education, 76*(4), 326-343.

Walkerdine, V. (1992). Progressive pedagogy and political struggle. In C. Luke & J. Gore (Eds.), *Feminisms and critical pedagogy* (pp. 15-24). New York: Routledge.

Walkerdine, V. (1997). Difference, cognition, and mathematics education. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 201-214). Albany: SUNY.

Wenglinsky, H. (1997). *When money matters: How educational expenditures improve student performance and how they don't*. Princeton, NJ: Educational Testing Service.

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Educational Policy Analysis Archives, 10*(12), 31.

Wenglinsky, H. (2004). Closing the racial achievement gap: The role of reforming instructional practices. *Education Policy Analysis Archives, 12*(64), 22.

Wexler, P. (1988). Symbolic economy of identity and denial of labor: Studies in high school number 1. In L. Weis (Ed.), *Class, Race, and Gender in American Education* (pp. 302-315). Albany: State University of New York.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*, 461-481.

Whittington, D. (2002). *2000 National survey of science and mathematics education: Status of middle school mathematics teaching* (report). Chapel Hill, NC: Horizon Research, Inc.

Willis, P. (1981). Elements of a culture. In *Learning to Labor: How working class kids get working class jobs*. New York: Columbia University Press.

Willis, S., & Brophy, J. (1974). The origins of teachers' attitudes towards young children. *Journal of Educational Psychology, 66*(4), 520-529.

Willms, J. D. (2006). *Learning divides: Ten policy questions about the performance and equity of schools and schooling systems* (report No. UIS/WP/06-02). Montreal: UNESCO Institute for Statistics.

Wright, C. (2006, February 19, 2006). Neighborhood schools promote segregation. *Chapel Hill News,* pp. A1, A10.

Xin, T., Xu, Z., & Tatsuoka, K. (2004). Linkage between teacher quality, student achievement, and cognitive skills: A rule-space model. *Studies in Educational Evaluation, 30*, 205-223.

Zaslavsky, C. (1997). World cultures in the mathematics class. In A. B. Powell & M. Frankenstein (Eds.), *Ethnomathematics: Challenging Eurocentrism in mathematics education* (pp. 307-320). Albany: SUNY.

Zaslavsky, C. (1999). *Africa Counts* (3rd ed.). Chicago: Chicago Review Press.

Zevenbergen, R. (2000). Cracking the code of mathematics classrooms: School success as a function of linguistic, social, and cultural background. In J. Boaler (Ed.), *Multiple*

perspectives on mathematics teaching and learning (pp. 201-223). Westport, CT: Ablex.